이학박사 학위논문

Informatics techniques to navigate transcriptome space with biological networks from gene to pathway to phenotypes

생물학적 네트워크를 이용하여 유전자로부터 패스웨이, 표현형까지의 전사체 공간을 탐색하는 정보학 기법

2019년 8월

서울대학교 대학원

협동과정 생물정보학

문 지 환

이학박사 학위논문

Informatics techniques to navigate transcriptome space with biological networks from gene to pathway to phenotypes

생물학적 네트워크를 이용하여 유전자로부터 패스웨이, 표현형까지의 전사체 공간을 탐색하는 정보학 기법

2019년 8월

서울대학교 대학원

협동과정 생물정보학

문 지 환

Informatics techniques to navigate transcriptome space with biological networks from gene to pathway to phenotypes

생물학적 네트워크를 이용하여 유전자로부터 패스웨이, 표현형까지의 전사체 공간을 탐색하는 정보학 기법

지도교수 김 선

이 논문을 이학박사 학위논문으로 제출함

2019 년 5 월

서울대학교 대학원

협동과정 생물정보학

문 지 환

김민수의 이학박사 학위논문을 인준함

2019 년 6 월

위 원 장	황대희
부위원장	김선
위 원	한원식
위 원	이슬
위 원	채희준

Abstract

Informatics techniques to navigate transcriptome space with biological networks from gene to pathway to phenotypes

Ji Hwan Moon
Interdisciplinary Program in Bioinformatics
College of Natural Sciences
Seoul National University

Transcriptome data, genome-wide measurement of transcripts, has been used to increase our understandings of biological processes at transcription level significantly. Analysis of transcriptome data involves a series of steps from identification of differentially expressed genes (DEGs) to pathway enrichment analysis to association with phenotypes. There exist several hurdles at each step that need to be addressed with state of the art bioinformatics techniques. For example, the complex nature of living organisms can be represented as a network where the nodes are the interacting entities such as genes or pathways and the edges are the interactions between the nodes. Network analysis is crucial in that it can reveal the hidden associations between transcriptome data and phenotypes. In addition, network propagation has emerged as a technique

to measure the influential power of nodes in a network. Network propagation has demonstrated its utility on biological context by many studies and has been contributing to invaluable discoveries in biological and medical science fields. In my doctoral study, I explored and analyzed trasncriptome at various levels using machine learning, network information and network propagation techniques.

My thesis consists of three studies. The first study was to develop an accurate and stable method for determining differentially expressed genes using machine learning techniques. The second study was to develop a novel method to investigate interactions among biological pathways using explicit gene expression information from RNA-seq. The last study was to perform analysis of xenotransplant transcriptome data using various methods including the network propagation technique.

In the first study, MLDEG, a machine learning approach to identify DEGs using network property and network propagation, was developed. Currently available DEG detection methods have widely been used and contributed to new biological discoveries. Most of the methods use their own models to define DEGs. However, because the traits of transcriptome data vary significantly depending on the experimental designs and sequencing technologies, a single model can hardly fit all transcriptome data of different traits. In addition, setting cutoff values of p-values and fold change is arbitrary. Thus, the results yielded by the methods are often inconsistent and heterogeneous. MLDEG addresses these issues by building a model that uses network information and network propagation results as features. The goal of MLDEG is to train a model by using network-based features extracted from more likely true and false DEGs and use the model to classify DEGs from the genes that cannot be clearly defined as DEGs by existing methods. Tested on 10 high-throughput RNA-seq data, MLDEG showed better performances than the competing methods.

In the second study, I developed a Pathway INTeraction network construction method (PINTnet) that can construct a condition-specific pathway interaction network by computing shortest paths on protein-protein interaction (PPI) networks. Because pathways usually function in a coordinated and cooperative fashion, understanding interactions, or crosstalks, between pathways becomes as important as identifying perturbed single pathway. However, existing methods do not take into account the topological features, treating the pathways just as a set of genes. To solve the problem, PINTnet computes shortest paths on PPI networks mapped to each pair of pathways and creates subnetworks using the shortest paths. It then measures the activation status of pathway interaction using the product of closeness centrality and explicit gene expression quantity. The performance of PINTnet was evaluated using three high-throughput RNA-seq data and successfully reproduced the findings in the original papers of the data.

In the last study, I participated in a xenotransplantation study to elucidate the cause of chronic phase islet graft loss. Clinical islet transplantation is one of the promising options for type 1 diabetes but long-term outcome of graft function is not yet satisfactory. To reveal the mechanism of the graft loss in chronic phase, I carried out pathway interaction network analysis using PINTnet on a time-series porcine islet-transplanted rhesus monkey RNA-seq data and identified the activation of T cell receptor signaling pathway. The analysis results were supported by the biopsy result of liver sample that $CD3^+$ T cell heavily infiltrated the porcine islet. Additionally, I carried out gene prioritization using network propagation to verify five graft loss-relevant scenarios. The result suggested that T cell-mediated long-term graft loss was the most probable scenario.

In summary, my doctoral study used network information, network property,

and network propagation to identify DEGs and predict pathway interactions. In addition, I participated in a xenotransplantation research and carried out pathway interaction network analysis and network propagation to reveal the possible cause of chronic phase islet graft loss. Utilizing network information and network propagation was very effective to discover the relationships among biological entities and analyze the complex phenotypes.

Keywords: protein-protein interaction, shortest path, network propagation, differentially expressed gene, xenotransplantation, chronic phase islet graft loss
Student Number: 2012-30906

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Transcriptome refers to the complete set of transcripts and their quantity in a cell for a specific condition. Transcriptome analyses are essential in biological researches because the cellular states can be identified by such analyses (Mortazavi *et al.*, 2008). There are two major technologies measuring transcriptome: microarray and sequencing. Due to the greater information content and the lower cost, the sequencing technology is more preferred than microarrays. The transcriptome data have been widely used to reveal the biological mechanisms underlying many phenotypes or diseases. To analyze transcriptome data of more than 20,000 genes, various computational techniques have been developed and used. Differential expression analysis and gene set enrichment analysis are the most widely used analysis methods. Although these methods are successfully used to discover insights on new biological knowledge, they do not consider complex interactions among genes. Networks are most natural and effective tools to represent and model interactions among genes. Indeed, there exist multiple types of biological networks and the networks share common fea-

tures; a network is a set of interactions or relations between different entities. The entities can be any of biological elements and the interactions can represent either positive relation or negative relation according to the types of biological elements. The interactions are determined by biological experiments or inferred by computational methods, and stored in various resources such as STRING database (Szklarczyk *et al.*, 2016), a repository of protein-protein interactions of different species, or KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa and Goto, 2000). The resources are very useful and have been used in studying biological mechanisms.

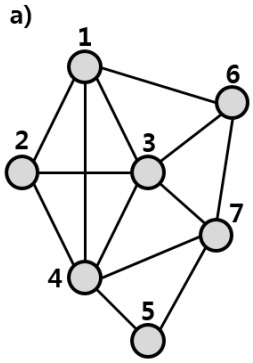
In this thesis, I investigated three research problems ranging from gene-level analysis to pathway-level analysis using network information and network propagation. The first problem is to determine differentially expressed gene (DEG) more accurately using biological networks. The second problem is to infer condition-specific pathway interaction network from transcriptome data. The last problem is about the study to infer the cause of graft loss in xeno-transplantation.

1.1 Background

In my doctoral study, I used two computational techniques, networks and machine learning. Thus, in this section, I explain technical background on the two techniques.

1.1.1 An introduction to network theory and its application to the fields of biology

A network is a set of nodes and edges. It can be denoted as $G = (V, E)$ where V is a set of nodes and E is a set of edges where each edge is a pair of nodes in V. It is called a directed network if there is a direction in each edge that comes inside



b)

	1	2	3	4	5	6	7
1	0	1	1	1	0	1	0
2	1	0	1	1	0	0	0
3	1	1	0	1	0	1	1
4	1	1	1	0	1	0	1
5	0	0	0	1	0	0	1
6	1	0	1	0	0	0	1
7	0	0	1	1	1	1	0

Figure 1.1: An example of an undirected network

or outside of nodes. In a directed network, an edge is an ordered pair (v_i, v_j) that has a direction from v_i to v_j where i and j are the indices of the nodes smaller than or the same as the total number of the nodes in V. In an undirected network, an edge is just a set of any two nodes $\{v_i, v_j\}$ in V. For example, as shown in Figure 1.1 a), there is an undirected network where $V = \{1, 2, 3, 4, 5, 6, 7\}$ and $E = \{\{1, 2\}, \{1, 3\}, \{1, 4\}, \{1, 6\}, \{2, 3\}, \{2, 4\}, \{3, 4\}, \{3, 6\}, \{3, 7\}, \{4, 5\}, \{4, 7\}$ $, \{5, 7\}, \{6, 7\}\}$. An adjacency matrix is a square matrix used to represent a network. The elements indicate if any pair of nodes in a network are connected. The equation below shows an adjacency matrix of an undirected and unweighted network.

$$A_{i,j} = \begin{cases} 1 & \text{if } \{v_i, v_j\} \in E \\ 0 & \text{if } \{v_i, v_j\} \notin E \end{cases} \tag{1.1}$$

Figure 1.1 b) shows the adjacency matrix of a network in Figure 1.1 a). In addition, it is called a weighted graph if each edge has a weight. The degree of a node is the number of edges incident on the node. It is denoted as $deg(v_i)$ and, for example, the degree of node 1 in the network in Figure 1.1 a) is $deg(1) = 4$.

There exist different types of biological networks where biological entities

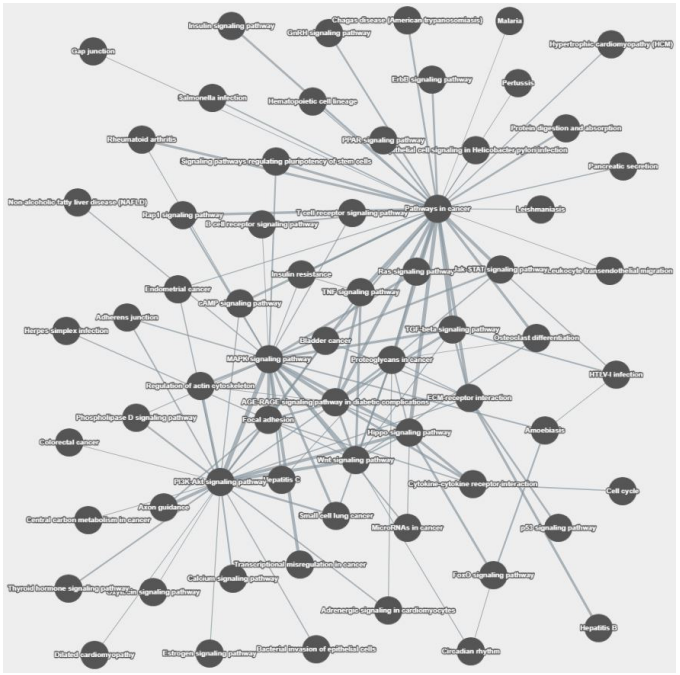


Figure 1.2: An example of biological network

are the nodes and their interactions are represented as edges. For example, transcriptional regulatory networks represent genetic regulatory relationships in cells. The nodes are transcription factors (TFs) and genes and the edges are between TFs and the target genes. Metabolic networks reflect a set of biochemical reactions. The nodes are metabolites and the edges are biochemical reactions between the metabolites. Once constructed, the biological networks are analyzed using various computational methods.

The simplest approach might be to search the direct neighbor genes of the genes related to a certain disease (Oti *et al.*, 2006). However, there will be false positive results when irrelevant genes are directly connected to the disease genes or relevant genes have no direct interaction with the disease genes. One of the alternative approaches is to use distance-based methods (Franke *et al.*,

2006). The distances are measured using shortest path computation. Network propagation is a path-based method that can consider all paths simultaneously and rank the genes. It becomes popular in gene prioritization and prediction (Cowen *et al.*, 2017).

1.1.2 An introduction to machine learning

Machine Learning is a field of artificial intelligence for developing algorithms and techniques to enable computers to learn (Simeone *et al.*, 2018). Machine learning processes are similar to data mining and predictive modeling. It searches patterns in data and update its actions according to the patterns. There are two major categories of machine learning. One is supervised learning and the other is unsupervised learning. Supervised learning requires a guideline about inputs and desirable outputs (Kotsiantis *et al.*, 2007). When the training is complete, the trained model is used on new data. Classification is one of the examples of supervised learning. Classification of real data is very difficult because, in many cases, it is in a gray zone that it is hard to decide which class the data belong to. There are several studies using classification in the fields of biology. For example, there is a study to classify samples using gene expression as features (Wu, 2005). Another example is to classify original breast cancer tissue in a patient-derived tumor xenograft model (Wang *et al.*, 2018). Unsupervised learning infers patterns from data without any information about desirable outputs (Sathya and Abraham, 2013). It uses iterative approach to discover the underlying structure of the data. Clustering is one of the examples of unsupervised learning. In this thesis, classification is used in an ensemble machine learning method that integrates the advantages of existing methods and classifies DEGs using transcriptome data.

1.2 Three problems in my doctoral study

1.2.1 Problem 1: DEG detection

DEGs are the genes of which the expression changes or differences are observable between two experimental conditions. The identification of DEGs is very important in transcriptome analysis because it can lead to a new discovery associated with the conditions. There exist multiple methods to detect DEGs. The DEG detection tools have been widely used to analyze transcriptome data and to characterize biological mechanisms underlying phenotypes, e.g., human diseases. Although the simplest way to identify DEGs is to calculate the fold change of expression of each gene, it is not welcomed because the results can be biased by the huge fold changes obtained by the comparison of small expression values or can be different depending on the threshold. Instead, statistical model is adopted for more robust DEG identification. The common assumption of the statistical models is that gene expression is sampled from a statistical distribution. Null hypothesis is that two sets of gene expression are sampled from the same distribution and alternative hypothesis is that two sets of gene expression are sampled from two different distribution. Differential expression is more likely when the two distributions less overlap. The overlap is determined by means and standard deviations of two distributions. Each DEG detection method uses its own model to define DEGs. For example, EBSeq estimates the posterior likelihoods of differential and equal expression by the aid of empirical Bayesian methods, assuming negative binomial distribution (Leng *et al.*, 2013). edgeR determines differential expression using empirical Bayes estimation and exact test based on a negative binomial model and the Trimmed Mean of M values (TMM) normalization procedure is carried out to account for the different sequencing depths (Robinson *et al.*, 2010). DESeq2 uses a negative

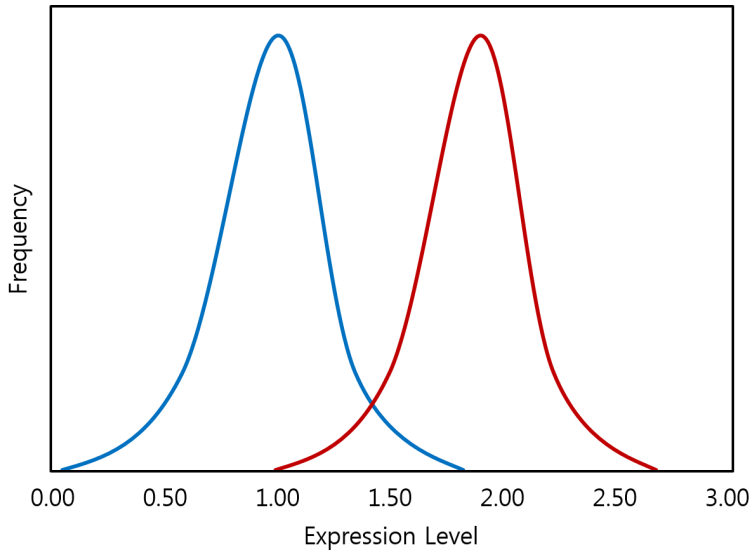


Figure 1.3: The concept of statistical model in DEG detection

binomial model similar to edgeR. When estimating dispersion, it models the observed relationship between the mean and variance and data-driven parameter is estimated (Love *et al.*, 2014). Limma is based on linear modeling. It is recommended to use TMM normalization of the edgeR package (Ritchie *et al.*, 2015).

Challenge: robust detection of DEGs There are some limitations of the existing methods using statistical models. The existing methods usually a single model to assume the distribution that the data follow. However, the traits of transcriptome data are different regarding various factors such as experimental conditions and technology used to measure the gene expression level. A single model cannot fit all the traits. As a result, the differential expression results by the methods are heterogeneous as shown in Figure 1.4. It is required to

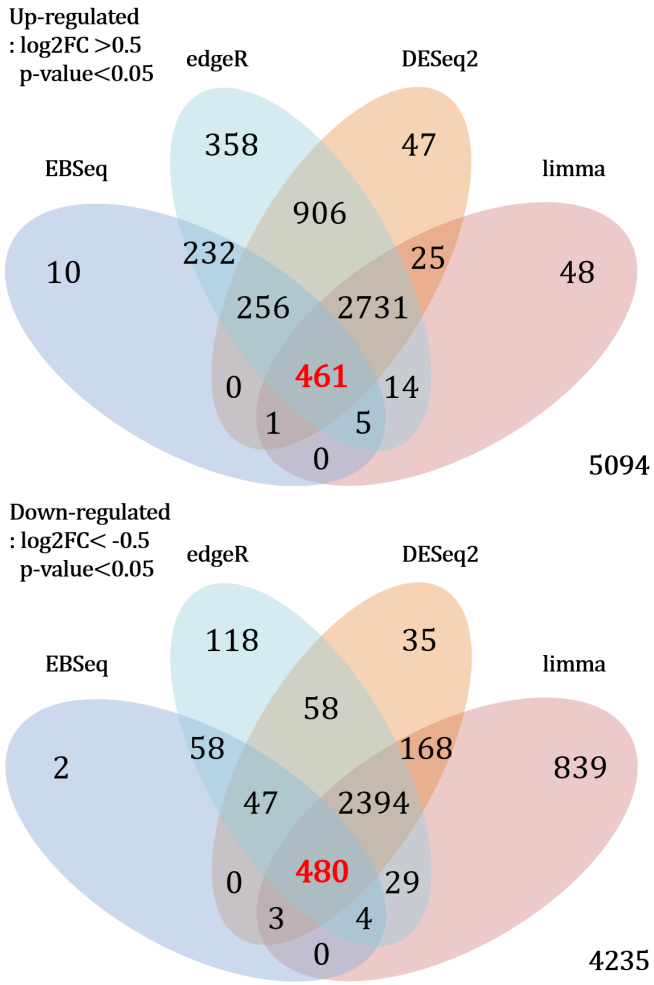


Figure 1.4: The heterogeneous results of four different DEG detection methods on TCGA breast cancer data

1.2.2 Problem 2: Pathway interaction analysis

A biological pathway is a set of interactions that sequentially occur in a cell. It leads to a certain cellular product or change. There are many biological pathways but they are originally in one big network. It is dissected into biological pathways by researchers. Perturbation of pathways directly affects the

cellular state. It usually originates in disease conditions so the identification of perturbed pathways is very important in revealing dysregulated biological mechanisms (Ramanan *et al.*, 2012). Gen Set Enrichment Analysis (GSEA) (Subramanian *et al.*, 2005) is one of the examples of perturbed pathway identification method. It tests the significance of the ratio of DEGs in each pathways to determine perturbed pathways. However, GSEA and methods similar to GSEA have a limitation that the methods merely consider pathways a set of genes. Thus, the network information is neglected. Meanwhile, there are some methods utilizing network information for the task such as signaling impact analysis (SPIA) (Tarca *et al.*, 2009). The methods are based on the concept that if the same number of genes are differentially expressed both in pathway A and pathway B, then the pathway that the DEGs are in a relation of interactions is biologically more meaningful.

Challenge: pathway interaction analysis Pathways usually function cooperatively. Identification of the interacting perturbed pathways is as important as that of perturbed individual pathways. There are several approaches to find the interactions of perturbed pathways. Most of the approaches are based on the shared components of pathways. For example, the simplest one is to consider the shared components such as genes or proteins between pathways. It assumes that shared genes may mediate interactions and predicts such interactions by testing the significance of the overlapping genes between pathways using hypergeometric test such as Fisher's exact test (Francesconi *et al.*, 2008). Another approach is to estimate interactions using protein-protein interaction (PPI) information. This approach assumes that any two interacting pathways may have more edges connected in PPI than expected. Due to the insufficient information provided by the overlapping genes, the identification of pathway

interaction may result in false positive. In fact, there is a study that the effect of overlapping genes between pathways causes false positive results in pathway analysis (Donato *et al.*, 2013). Therefore, it is required to consider not only the overlapping genes but also the interacting genes.

1.2.3 Problem 3: Analysis of transcriptome from pig-to-nonhuman primate islet xenotransplantation

Islet replacement can be a preferable option for long-term diabetic patients receiving insulin therapy. However, it is limited by the number of islet donors. Due to the organ shortage, islet xenotransplantation becomes a promising treatment option for the patients who suffer from type 1 diabetes. Among other species, pigs are suited to the purpose the best for the following reasons. First of all, pig insulin is very similar to that of human. Second of all, pig islets are not easily damaged by recurrent type 1 diabetes autoimmunity. Thirdly, amyloid is not accumulated by pig islets. Lastly, pigs can be genetically modified (Marigliano *et al.*, 2011). In fact, specific pathogen-free transgenic miniature pigs are bred in Biomedical Center for Animal Resource Development at Seoul National University and used as islet donors for xenotransplantation researches. The experimental results of islets outperformed that of whole organ and it shows the availability of islets. In addition, hyperacute rejection rate is very low and short-term results of islet function after transplantation are improved. A recent study showed immunosuppression was effective for preserving islet mass and controlling diabetes in pig-to-nonhuman primate models and normoglycemia was maintained over six months in four out of five monkeys (Shin *et al.*, 2015). However, long-term results are not satisfactory and the cause of long-term graft loss is not yet discovered.

Challenge: early detection of long-term graft loss and determination of the cause Once transplanted, the function of graft is monitored using peripheral blood. The Enzyme-Linked ImmunoSpot (ELISpot) assay to measure the frequency of cytokine secretion and intravenous glucose tolerance test (IVGTT) to estimate insulin sensitivity are the examples. However, the symptoms detected by such methods are merely signs that are observable after the onset of rejection. Gene-level or pathway-level early detection is not possible through such methods. Moreover, it is difficult to determine what has driven the rejection. There are several hypotheses on how long-term graft loss is driven. The hypotheses include ER stress (Fonseca *et al.*, 2011), lipotoxicity (Lee *et al.*, 2007), islet exhaustion (Kim and Yoon, 2011), long-term graft rejection and toxicity of immunosuppressant (Barlow *et al.*, 2013). Each hypothesis holds its own key genes evaluated by related studies (Chen *et al.*, 2013). It is important to measure the global effect of the genes and associate the effect with transcriptome data or phenotypes for determining the cause of rejection.

1.3 My network-based approaches to three research problems

In the previous sections, gene-level and pathway-level analysis methods and their limitations are introduced. My doctoral study includes network-based approaches to address the challenges. Additionally, it includes a real-world study that my methods were used. The details are as follows.

1. **MLDEG: A machine learning approach to identify differentially expressed genes using network property and network propagation:** Most of the existing differential analysis methods use their own model. However, a single model cannot fit all traits of different data so

the results on the same data are heterogeneous. MLDEG integrates the existing methods and builds a model using network-based feature to solve the problem and suggest a credible set of DEGs.

2. **PINTnet: Construction of condition-specific pathway interaction network by computing shortest paths on weighted PPI:** A biological pathway is a set of genes and their interactions. To identify the interactions between pathways, it is not enough to consider only overlapping genes of pathways. PINTnet considers neighboring genes of overlapping genes and computes shortest paths from one pathway to the other pathway on a weighted PPI subnetwork for every pair of pathways. Then it calculates activation score of the pathway interaction using gene expression data. The activated interactions are used as edges to construct a condition-specific pathway interaction network.

3. **Bioinformatics analyses with peripheral blood RNA-sequencing unveiled the cause of the graft loss after pig-to-nonhuman primate islet xenotransplantation model:** Pathway interaction analysis introduces an approach that can find the factors affecting long-term graft rejection in early stage. In addition, network propagation shows a possibility to investigate the system-wise landscape of whole genes in association with long-term graft loss-related genes.

1.4 Outline of thesis

Chapters 2, 3 and 4 introduce independent studies related to differential expression analysis, pathway interaction network construction and xenotransplantation, respectively, from gene-level study to pathway-level study. Each of the studies uses network information and network-based algorithms to solve the

problems. In Chapter 2, a machine learning approach identifying DEGs using network information and network propagation is described. It aims to train a model using network-based features extracted from more likely true DEGs and classify the genes that cannot be clearly defined by DEGs by the existing methods. Chapter 3 describes PINTnet, a method to construct a condition-specific pathway interaction network by computing shortest paths on a weighted PPI network. Chapter 4 presents a xenotransplantation study to elucidate the cause of long-term graft loss. Chapter 5 summarizes the studies. The bibliography of the cited references is at the end of the thesis.

Chapter 2

A machine learning approach to identify differentially expressed genes using network property and network propagation

A pivotal step in transcriptomic analysis is to identify genes of which expression changes are significant and associated with the phenotypes or experimental conditions. Thus, a number of tools have been developed for identifying differentially expressed genes (DEGs) in transcriptome data. These tools have been used extensively for numerous research projects, contributing to discoveries of new biological mechanisms. However, no single statistical or machine learning model for DEG detection can perform consistently well for datasets of different traits and the performances of the existing DEG methods vary a lot for datasets measured under different conditions. In addition, setting a cutoff value for the significance of differential expressions is one of the confounding factors to determine DEGs.

I address these problems by developing a machine learning method. My method first prepares training data by compiling DEG predictions by existing methods and builds a machine learning model using network features to model gene-gene interactions and features for the influence of a gene on other genes by network propagation techniques. Then, DEG candidates, the genes that are predicted with weak evidences by the existing tools, are classified by the machine learning model. Tested on 10 RNA-seq datasets, my method showed an excellent performance; my method won the first place in detecting ground truth (GT) genes in eight datasets and could find almost all GT genes in six datasets. On the other hand, the performances of the compared methods varied significantly for the 10 datasets. By design, my method can accommodate any new DEG method to improve the performance. The source code of my method is available at https://github.com/jihmoon/MLDEG.

2.1 Background of differential expression analysis methods

It is one of the pivotal steps in transcriptomic analysis to identify genes of which expression changes are significant and associated with the phenotypes or experimental conditions. Most of the existing methods are based on the comparisons of gene expression levels among multiple conditions. The gene expression comparisons range from a simple implementation of comparing logarithmic fold change to statistical hypothesis testing. However, input to the statistical testing varies a lot depending on many factors such as the experimental conditions, short read mapping tools used, and data normalization methods. Because of these confounding factors, the selection of a proper DEG detection method is not straightforward.

In many cases, biologically significant phenomena are the results of cooperation and interaction of multiple genes (Wang *et al.*, 2008). While the effects of individual genes may be small and even ignorable, considering the genes in pathways or the network can explain complex phenotypes or diseases. Thus, there have been a lot of studies to reveal the regulatory relations among genes for decades. For example, STRING is a protein-protein interaction database storing over nine million relations among genes at protein level from more than 2,000 organisms (Szklarczyk *et al.*, 2016). Kyoto Encyclopedia of Genes and Genomes (KEGG) contains a collection of manually drawn pathways (Kanehisa and Goto, 2000). There are evidences that utilizing the network information will result in more interpretable sets of statistically significant genes. For example, Gene Set Enrichment Analysis (GSEA) (Subramanian *et al.*, 2005), Parametric Analysis of Gene Set Enrichment (PAGE) (Kim and Volsky, 2005) and Generally Applicable Gene-set/Pathway Analysis (GAGE) (Luo *et al.*, 2009) utilized network and pathway information and achieved good performance in detecting phenotypically related gene sets. Tarca *et al.* 2008 and Nam *et al.* 2014 directly incorporated network structure and take into account the subnetwork of DEGs to search pathways relevant to the phenotypes but these methods are not designed to search for DEGs.

There exist only a few methods that utilize network information to identify DEGs. pathDESeq (Dona *et al.*, 2017) is an example of such methods. It searches for DEGs using a three-state Markov Random Field model that promotes genes within the same pathway to have the same differential expression status. Network propagation is one of the effective ways to consider and amplify the relationship among genes (Cowen *et al.*, 2017). It is based on diffusion process of influence through a network and can measure the global similarity of nodes. It assumes that genes associated with the same phenotypes tend to

interact. Thus a network propagation analysis produces a list of genes that are influenced by their direct neighbors. It is widely used in gene prioritization, drug-target association and even DEG detection. Signed-NP (Zhang *et al.*, 2012) is one of the examples to utilize the network propagation technology to detect and classify DEGs and copy number variations (CNVs).

2.1.1 Motivation

There are several issues with the current DEG selection methods. First, traits of the transcriptome data vary a lot depending on the experimental designs and the technologies used to measure transcriptome data. Thus, *a single model cannot fit all transcriptome datasets of different traits.* Additionally, setting cutoff values of p-value or fold change is a common practice to determine DEGs. However, setting a cutoff value for DEGs is arbitrary. *Depending on the cutoff values, the selection of DEGs varies significantly.* Moreover, it is possible that the expression changes are small but statistically significant. In addition, a high fold change, in some cases, may be the result of comparison between lowly expressed genes. Lastly, *most of existing DEG detection methods do not consider interaction among genes.* As a result, the performances of a method could be inconsistent for different datasets. More seriously, DEG selections by different methods differ widely for the same dataset as shown in Results and Discussion (Section 2.3.3). This is the main motivation for developing my machine learning method.

2.1.2 My machine learning approach

In this paper, I propose a machine learning-based DEG detection method, MLDEG, to address the issues of *data trait variance*, *arbitrary cutoff setting* and *lack of gene interaction information use*. The main idea is as follows. The

existing DEG methods are shown to be useful. Thus, my method utilizes, combines and refines the results of the existing methods using a machine learning approach. To train my model, I need positive and negative data. Setting a stringent cutoff for each method, I can easily get more likely true DEGs, especially when multiple DEG tools agree which genes are DEGs. Thus, I can obtain the positive data, i.e, true DEGs. The negative data can also be obtained by aggregating DEG predictions by the existing methods. More specifically, MLDEG combines and sorts DEG predictions by the existing methods, and then the genes with lower rankings are those that a majority or all of the DEG tools predicted as non-DEGs. The remaining genes that are not included in the training data are in the gray area. After training MLDEG with the positive and negative data, the goal is to classify which of the remaining genes in the gray area are DEGs. This way, I do not have to worry about setting an appropriate cutoff value for DEGs. To consider gene interaction information, MLDEG maps genes to a protein-protein interaction (PPI) network, calculates the network properties of each gene and carries out network propagation to measure the influence of each gene on the network. In short, my method selects the most probable DEGs, extracts features regarding gene interaction and influence among genes and then a machine learning model is built to classify genes in the gray area. The overview of my method is shown in Figure 2.1.

2.2 Methods

In this section, I explain how to obtain the training data, how to extract features for DEG prediction and how to train MLDEG.

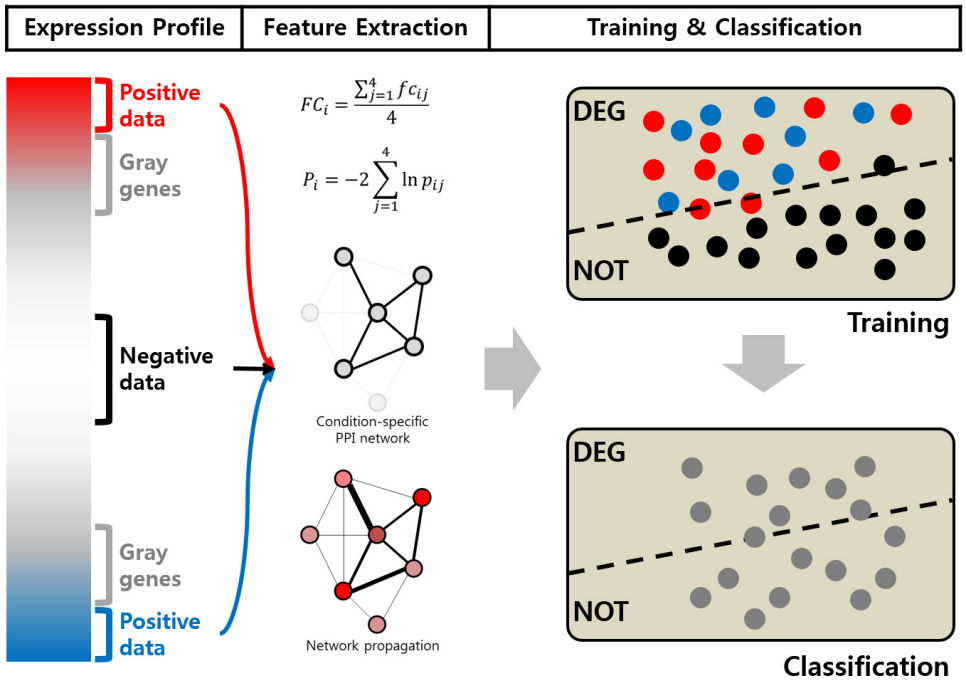


Figure 2.1: The overview of MLDEG

2.2.1 Training and Test Data

The first step of my method is to prepare training data. In order to do so, my method uses four different DEG methods to generate candidate DEGs. The methods used are EBSeq, DESeq2, edgeR, and limma. After executing the methods on input gene expression profiles, the results are combined and genes are sorted in the order of significance of the gene expression changes. P-values of the methods for each gene are combined and converted to a single score using Fisher's method (Fisher, 1932). The equation is as follows:

$$P_i = -2\sum_{j=1}^{4} \ln p_{ij} \tag{2.1}$$

where i indicates a gene and j indicates each DEG tool. P_i is used as the criterion of significance. In addition, the average of the log of fold change values

is used to measure the level of expression change.

$$FC_i = \frac{\sum_{j=1}^{4} fc_{ij}}{4} \tag{2.2}$$

In equation (2.2), i indicates a gene, j indicates each DEG tool, and fc_{ij} indicates the log2 of the expression fold change of gene i calculated by DEG tool j. It is obvious that a higher P_i and also a higher FC_i increase the probability that a gene is indeed differentially expressed. Therefore, by applying strict cutoff values to P_i and FC_i, my method selects positive data for training data. My method sorts the genes based on the combined p-value P_i in descending order and the genes of which the combined p-value is higher than the cutoff are considered as candidates of positive data. Among the candidate genes, the genes satisfying the fold change cutoff are finally selected as positive data. Genes with the low combined p-value are selected as negative data. The number of genes in the negative data is the same as the number of genes in the positive data, so that the learning can be done with a balanced data. Genes that are identified as DEGs by at least one DEG tool are considered candidate DEGs, or genes in the gray area. My method uses these candidate DEGs as test data.

2.2.2 Features

The features used in my method are three types. The first type is features that are related to gene expression. The expression feature consists of the average fold change FC_i and a combined p-value P_i where i indicates a gene as described in the previous section. The second type is network property features. The network property features deal with the degree of each gene and Pearson's correlation coefficient (PCC) of each edge between the genes on a network. The last type is network propagation feature which is the probability calculated by network propagation algorithm. Network property and network propagation play a pivotal role in my method. I expect that the features would reflect the

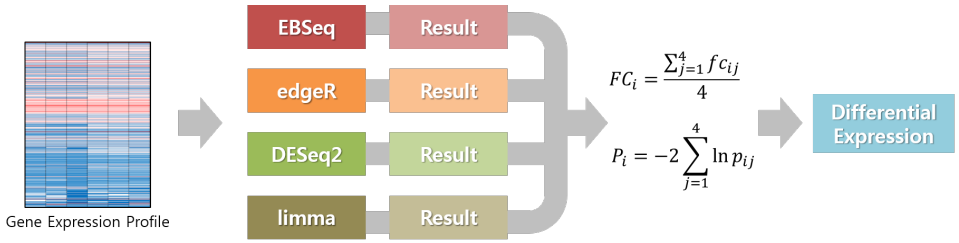


Figure 2.2: Differential expression feature extraction process

aspect that genes work in a cooperative and coordinated fashion by interacting each other. The features are shown in Table 2.1 and the overview of the feature extraction process is depicted in Figure 2.2 and Figure 2.3 for differential expression feature and network property and network propagation feature, respectively.

Table 2.1: List of Features

Feature Name	Feature Type
Average of log 2 fold change	Differential Expression
Combined p-value	Differential Expression
Condition-specific degree	Network Property
Ratio of condition-specific degree	Network Property
Mean of correlation coefficients	Network Property
Standard deviation of correlation coefficients	Network Property
Mean of correlation p-values	Network Property
Standard deviation of correlation p-values	Network Property
Probability by network propagation	Network Propagation

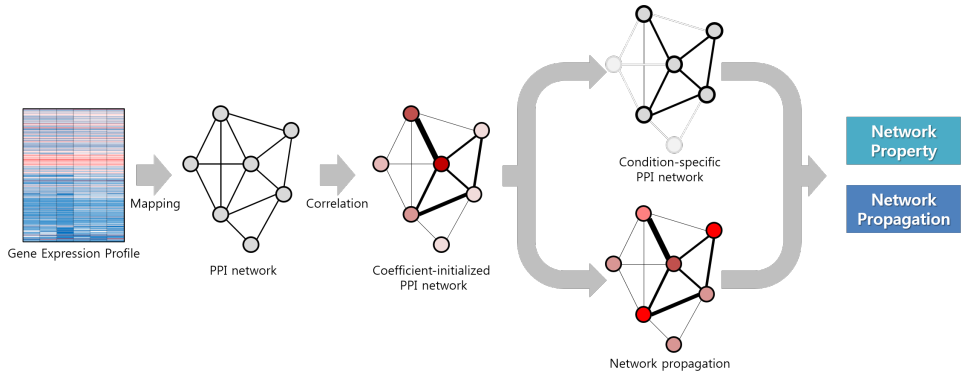


Figure 2.3: Network property and network propagation feature extraction process

2.2.3 Network Property

Genes are mapped to a PPI network downloaded from STRING database so that network property features can be calculated. Once the genes are mapped to a PPI network, my method calculates PCC of gene expression referring to **A**, the adjacency matrix of the PPI network. Then the PPI network is pruned by the criteria of PCC of 0.7 and a correlation p-value of 0.05. Any edge that has PCC less than 0.7 or p-value bigger than 0.05 is pruned and a condition-specific PPI network is constructed. "Condition-specific degree" and "Ratio of condition-specific degree" are the features coming from the condition-specific PPI network. For example, a gene has five direct neighbor genes so the degree of the gene is five. However, only two of the five edges have PCC larger than and p-value less than cutoff values so the other three edges are pruned. Then I can say the condition-specific degree of the gene is two. Dividing the condition-specific degree by the original degree of the gene, I can also say the ratio of condition-specific degree is 0.4. Basically, my method calculates PCC on the edges but the features are assigned to genes. If a gene belongs to multiple edges, then the

gene participates in the correlation calculation the number of the edges the gene belongs to and the gene has as many PCCs as the number of the edges. “Mean of correlation coefficients”, “Standard deviation of correlation coefficients”, “Mean of correlation p-values” and “Standard deviation of correlation p-values” are the features taking the issue into account and projecting the edge-level values to the node-level values. “Mean of correlation coefficients” is the average of PCC of edges that the gene belongs to. Likewise, “Standard deviation of correlation coefficients” is the standard deviation of PCC of edges that the gene belongs to. “Mean of correlation p-values” is the average of p-values for correlation of edges that the gene belongs to. Similarly, “Standard deviation of correlation p-values” is the standard deviation of p-values for correlation of edges that the gene belongs to.

2.2.4 Network Propagation

In addition to the network properties explained in the previous section, I need to measure how the genes with high influential possibility actually interact with and influence on other genes on a network. Considering direct neighbors, centrality calculation or shortest path computation can be used to achieve the goal but the time complexities of these schemes are dramatically increased depending on the size of network and the risk of false positives and false negatives is also high. To overcome these problems and quantify the association among the genes by measuring the global similarity of gene expression changes, network propagation is carried out and the results are used as network propagation feature. By doing so, MLDEG ranks the genes according to the influencing power. The first step of network propagation is initialization. I take the idea of Signed-NP and apply it to initialize the nodes in the PPI network. To carry out network propagation, the nodes and the edges must be initialized. To begin with, my

method creates a class vector $\mathbf{V} = \{v_1, v_2, \ldots, v_N\}$ where $v_k \in \{-1, 1\}$, k is the index of the samples and N is the total number of the samples. Either 1 or -1 is assigned to v_k if sample k is a treated sample or a control sample. Then, Pearson's correlation between the expression level of a gene and $\mathbf{V}$ is calculated to initialize the node. At this time, the absolute values of the coefficients are used. For the edge initialization, the PCC results calculated in the network property feature step are used. My method constructs a weighted PPI network $\mathbf{W}$ using the PCC results. Likewise, the weights of $\mathbf{W}$ are absolute values. When the initialization is finished, the network propagation is carried out. I used random walk with restart algorithm for network propagation. The equation is shown below:

$$\mathbf{p}^{t+1} = (1 - r)\mathbf{W}'\mathbf{p}^t + r\mathbf{p}^0 \tag{2.3}$$

where $\mathbf{W}'$ is column-normalized matrix of $\mathbf{W}$, t is a time step, $\mathbf{p}^0$ is the vector of initialized nodes, $\mathbf{p}^t$ is the vector at the current time step t, $\mathbf{p}^{t+1}$ is the vector at the next time step, and r is the restart rate. I use 0.7 for r and the algorithm stops the iteration when the $L1$ norm difference between $\mathbf{p}^t$ and $\mathbf{p}^{t+1}$ is smaller than 10^{-6}. When the network propagation is finished, $\mathbf{p}$ is assigned to genes as network propagation feature.

2.2.5 Machine Learning Algorithm

With the gene expression and network features, my method trains a model with the training data using WEKA (Witten *et al.*, 2016). Specifically, logistic regression is used as a machine learning algorithm. The trained model is validated by 10-fold cross-validation. After training, my model, MLDEG, classifies which genes in the gray area are DEGs. The identified DEGs are then sorted by the prediction confidence in descending order.

Table 2.2: Datasets used for evaluation

Name	Title	Accession Number
Dataset1	Acetylation-regulated interaction between p53 and SET reveals a widespread regulatory mode	GSE83635
Dataset2	Human Parvovirus Infection of Human Airway Epithelia Induces Pyroptotic Cell Death by Inhibiting Apoptosis	GSE102392
Dataset3	Alopecia areata is driven by cytotoxic T lymphocytes and is reversed by JAK inhibition	GSE45657
Dataset4	Macrophage Transcriptional Profile Identifies Lipid Catabolic Pathways That Can Be Therapeutically Targeted after Spinal Cord Injury	GSE84737
Dataset5	In Utero Caffeine Exposure Induces Transgenerational Effects on the Adult Heart	GSE79013
Dataset6	PI3K orchestration of the in vivo persistence of chimeric antigen receptor-modified T cells	GSE93386
Dataset7	Mutual epithelium-macrophage dependency in liver carcinogenesis mediated by ST18	GSE72403
Dataset8	Transcription factor Foxo1 is essential for IL-9 induction in T helper cells	GSE100634
Dataset9	Amino Acid Transporter Slc38a5 Controls Glucagon Receptor Inhibition-Induced Pancreatic α Cell Hyperplasia	GSE89636
Dataset10	Amino Acid Transporter Slc38a5 Controls Glucagon Receptor Inhibition-Induced Pancreatic α Cell Hyperplasia	GSE90116

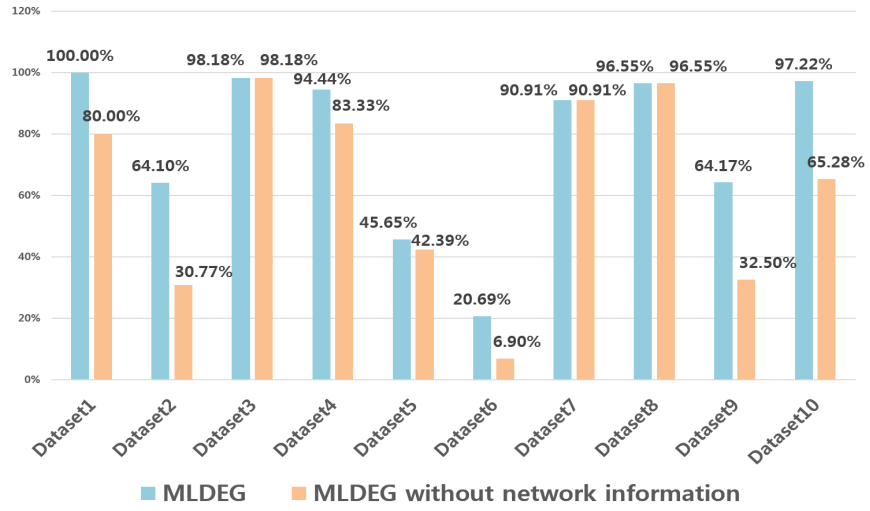


Figure 2.4: Comparison of the results of MLDEG with or without network information

2.3 Results and Discussion

I evaluated the performance of my method using 10 high-throughput RNA-seq datasets downloaded from GEO. The information of the datasets is shown in Table 2.2.

2.3.1 Experimental Data Description

I used the 10 datasets to evaluate the performance of my method in comparison with existing methods. The original papers of the datasets have their own phenotype-relevant gene sets verified by differential expression analyses. I considered the gene sets ground truth (GT) genes and evaluated how well my method could rescue the genes. Then, I compared the performance of my method to the selected four DEG tools.

The first dataset is from a study to identify the genes regulated by p53-SET

interplay (Wang *et al.*, 2016a). p53 C-terminal domain (CTD) acetylation is one of the early examples of non-histone protein acetylation and what it exactly does is undiscovered. The authors used a proteomic screen to reveal that the oncoprotein SET inhibited the transcriptional activity of p53 in unstressed cells but p53 was activated without the interaction with SET and tumor regression was observed in mouse xenograft models. They carried out RNA-seq analysis and reported 24 genes that were regulated by p53-SET interplay. Among the 24 genes, only 20 genes were mapped to PPI network. Thus, for this data set, the performance evaluation criterion was how many of the 20 genes were determined as DEGs.

The second dataset is generated by a study about human bocavirus 1 (HBoV1) infection (Deng *et al.*, 2017). It is a human parvovirus and a cause of acute respiratory tract infections in children. The authors showed that antiapoptotic proteins suppressing apoptosis and promoting pyroptosis were activated during HBoV1 infection using various methods including RNA-seq analysis. They presented 47 genes relevant apoptosis regulation during HBoV1 infection. 39 genes out of 47 genes were mapped to PPI network so 39 genes were used as the GT genes for the performance evaluation.

Alopecia areata (AA) is a autoimmune disease causing hair loss. It is mediated by T cells surrounding the hair follicle bulb. Dataset3 is from a study of AA (Xing *et al.*, 2014). The authors have previously identified a cytotoxic subset of $CD8^{+}NKG2D^{+}$ T cells in human AA hair follicles and the importance of two NKG2D ligands in pathogenesis. They used the C3H/HeJ mouse model to determine the contribution of $CD8^{+}NKG2D^{+}$ T cells to AA pathogenesis. Gene expression of flow sorted $CD8^{+}NKG2D^{+}$ T cells isolated from C3H/HeJ mice with alopecia were compared to that of $CD8^{+}NKG2D^{-}$ T cells from the same mice and 55 DEGs were identified including several NK-specific transcripts. I

set the 55 genes as GT genes and used the genes to evaluate my method.

It is well known that infiltrating macrophages are related to many pathological processes after spinal cord injury (SCI). However, the mechanisms how the macrophages function in accordance with the post injury are barely understood. To reveal the mechanisms, Zhu *et al.* 2017 obtained mRNAs that are directly related to macrophages from the injured spinal cord and sequenced using RNA-seq technology to characterize the gene expression profile and I used this data as dataset4. The data consists of three days post-injury and seven days post-injury. They reported lipid catabolism as the main biological process and canonical nuclear receptor pathway as their potential mediator. They also showed the relation between a lipoprotein, CD36, and recovery. In their study, seven days post-injury data best described the macrophages with lipid catabolism. The authors reported top DEGs pertaining to lipid metabolism or immune response. I used the 54 reported genes as GT genes.

Dataset5 is from a study showing the effects of caffeine to fetuses (Fang *et al.*, 2016). The authors exposed pregnant mice to caffeine at two embryonic stages and carried out analyses including cardiac gene expression assessment and RNA sequencing. They revealed that the timing of exposure was relevant to the long-term effects on cardiac function and morphology in that while adult mice exposed to caffeine from E6.5-9.5 showed abnormal cardiac function or morphology, adult mice exposed to caffeine from E10.5-13.5 were normal. They also reported 116 DEGs and showed that many cardiovascular disease pathways were significantly enriched. Among 116 genes, 92 genes were mapped to PPI network so I used 92 genes as ground truths.

The therapeutic effectiveness of chimeric antigen receptor (CAR)-modified T cells pertains to *in vivo* persistence but the reason of such persistence is not well known. Zheng *et al.*, 2018 used an acute myeloid leukemia (AML) model

to show the linkage between CAR expression and T cell differentiation. They found that the differentiation of CAR-T cells was modulated when treated with a PI3K inhibitor. RNA-seq analysis that the authors carried out revealed that PI3K/AKT pathway and glycolysis pathway as well as 30 genes belonging to the pathways were activated in CD33 CAR-T cells. I used their data as dataset6. Among 30 genes, 29 genes remained after gene ID conversion and the genes were used as GT genes for the performance evaluation.

Dataset7 is from a study pertaining to *ST18*. *ST18* is considered either tumor suppressor or oncogene in different human cancers. The exact role of *ST18* is not yet fully discovered. Ravà *et al.* 2017 tried to unravel that the role of *ST18* in tumor progression and maintenance using a mouse model with liver cancer. They demonstrated RNA-seq analysis of ST18-depleted tumors and control and confirmed inflammation-related genes such as NF-kB family, inflammatory cytokines and chemokines were down-regulated in ST18-depleted tumors. The genes control the recruitment and activation of myeloid cells to the tumor sites. The result suggests that *ST18* is the upstream regulator of the expression of genes related to inflammatory response guided by NF-kB. 11 genes related to the NF-kB pathway were reported and I used the genes as GT genes.

Dataset8 is from a study to show which transcription factor is crucial for interleukin 9 (IL-9) induction in IL-9-producing helper T (Th9) cells and Th17 cells (Malik *et al.*, 2017). IL-9 is closely related to allergic inflammation, autoimmunity, immunity to extracellular pathogens and anti-tumor immune responses so unraveling the IL-9 induction mechanisms is very important in understanding IL-9-mediated allergic inflammation and cancer immunotherapy. The authors analyzed the gene expression profile of Th9 cells and compared to that of Th0 cells by RNA-seq and identified forkhead family transcription factor Foxo1 was

highly ranked putative transcription factor as well as Th9-associated cytokines and Th9-associated transcription factors. From the results of further analyses, they confirmed that Foxo1 plays an important role for IL-9 induction in Th9 cells.

Dataset9 and Dataset10 are from the same study showing the role of sodium-coupled neutral amino acid transporter *Slc38a5* in regulating pancreatic α cell mass in mice (Kim *et al.*, 2017). Glucagon is important in maintaining blood glucose level. If glucagon signaling is interrupted by any factors such as genetic disruption, the amount of secreted glucagon is increased compared to normal state. This phenomenon is related to α cell mass increment. To elucidate what promotes α cell hyperplasia, the authors carried out RNA-seq analyses on pancreatic islets from glucagon receptor (GCGR)-blocking antibody-treated mice or GCGR knockout mice. From the results, they confirmed that *Slc38a5* was highly expressed and its deficiency with GPCR inhibition or knockout showed reduced α cell mass. They reported 120 and 72 genes for GPCR inhibition study and GPCR knockout study, respectively, and I used the genes as GT genes for the performance evaluation.

2.3.2 Performance of Network Information Features

I evaluated how much my method was improved when the network information was used. To do so, I trained models for each datasets with network-based features and without network-based features and compared the results. The comparison result is shown in Figure 2.4.

MLDEG showed better performance when using the network information as features. To be more specific, I looked into the training results of the models with network-based features. As shown in Table 2.5, it was observed that the weights of network-based features were tend to be higher than the weights of the other

Table 2.3: Training results of the datasets

The numbers in the parentheses are the number of correctly classified genes.

Name	# of positive data	# of negative data	Accuracy
Dataset1	182 (181)	182 (182)	99.7253%
Dataset2	173 (171)	173 (171)	98.8439%
Dataset3	549 (549)	549 (549)	100%
Dataset4	792 (788)	792 (791)	99.6843%
Dataset5	47 (47)	47 (46)	98.9362%
Dataset6	38 (35)	38 (37)	94.7368%
Dataset7	746 (745)	746 (745)	99.866%
Dataset8	4683 (4682)	4215 (4215)	99.9888%
Dataset9	17 (17)	17 (17)	100%
Dataset10	51 (50)	51 (51)	99.0196%

Table 2.4: Evaluation results

The best results of each dataset are highlighted in boldface. The numbers in the parentheses are the rankings of the methods.

Name	# of GT genes	MLDEG	EBSeq	edgeR	DESeq2	limma
Dataset1	20	**20** (1)	6 (4)	16 (3)	5 (5)	**20** (1)
Dataset2	39	**25** (1)	8 (4)	17 (2)	6 (5)	16 (3)
Dataset3	55	54 (2)	27 (4)	**55** (1)	53 (3)	10 (5)
Dataset4	54	**51** (1)	29 (5)	48 (3)	34 (4)	**51** (1)
Dataset5	92	42 (2)	21 (4)	**76** (1)	30 (3)	0 (5)
Dataset6	29	**6** (1)	3 (3)	5 (2)	1 (4)	0 (5)
Dataset7	11	**10** (1)	**10** (1)	**10** (1)	**10** (1)	2 (5)
Dataset8	29	**28** (1)	**28** (1)	**28** (1)	**28** (1)	**28** (1)
Dataset9	120	**77** (1)	9 (5)	**77** (1)	11 (4)	35 (3)
Dataset10	72	**70** (1)	38 (3)	69 (2)	32 (4)	2 (5)

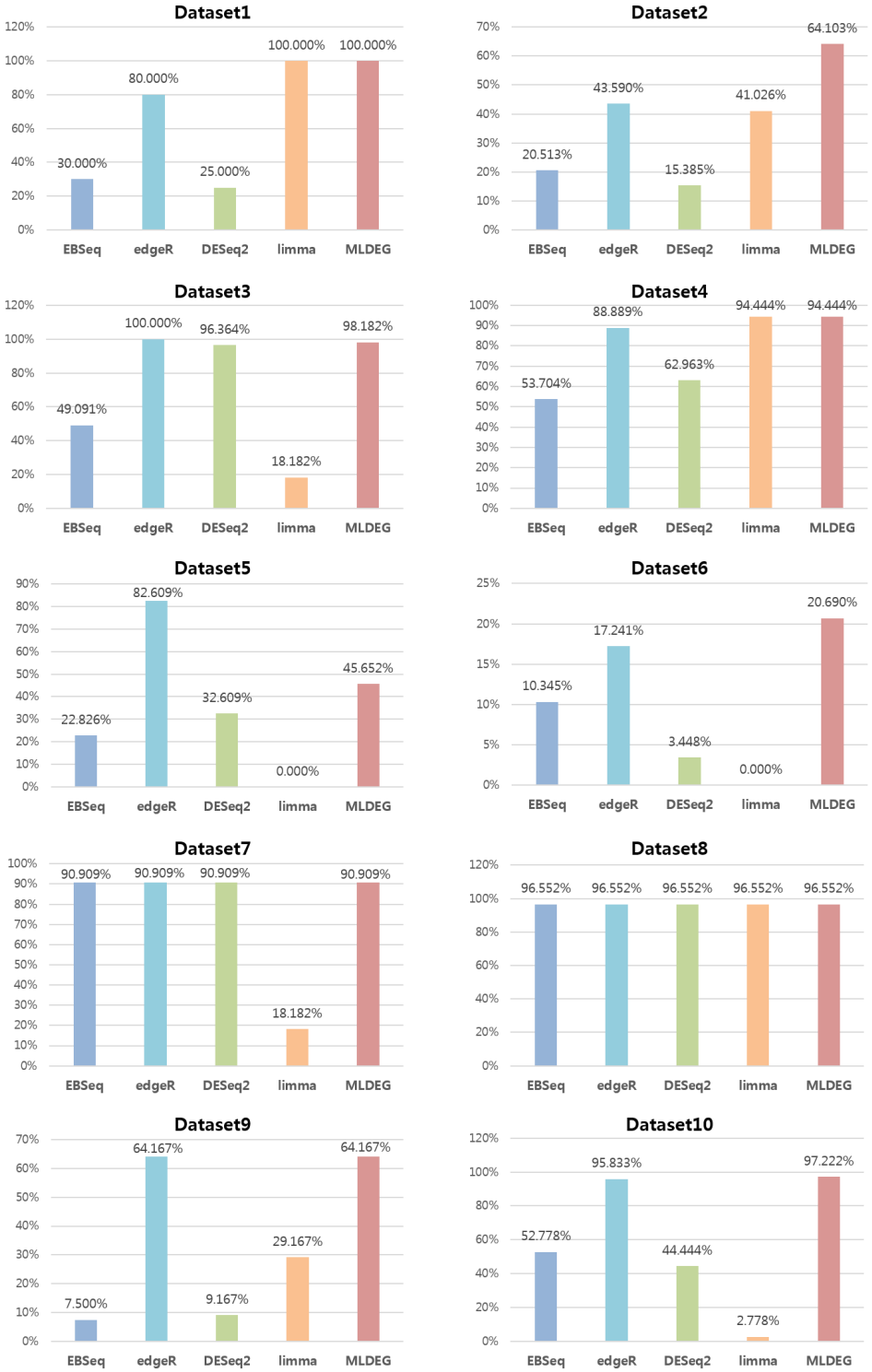


Figure 2.5: Performance comparison results

features. From the results, I concluded that considering network information was beneficial and useful in rescuing gray genes. Interestingly, MLDEG with network-based features and MLDEG without network-based features showed the same performance for dataset3, dataset7 and dataset8. Thus, I investigated the results of the three datasets and discovered that 50 out of 55 GT genes, nine out of 11 GT genes and 28 out of 29 GT genes were included in the positive data of dataset3, dataset7 and dataset8, respectively, so the potential of network information was not clearly observed.

Table 2.5: Odds ratio of features

The odds ratios that are bigger than 1.00E+05 are highlighted in boldface. Network-based features show higher odds ratios than differential expression features.

Name	log2FC	CombP	Degree	DegreeR	meanCC	stdCC	meanP	stdP	NP
Dataset1	**8.98E+21**	1.30E+00	1.79E+00	**Infinity**	**1.88E+205**	**2.37E+301**	0	0	**Infinity**
Dataset2	2.62E+01	1.17E+00	8.80E-01	**3.29E+179**	0	**1.72E+64**	0	0	**Infinity**
Dataset3	8.66E-01	2.35E+00	1.00E+00	8.70E-03	4.36E-01	1.50E+03	8.90E-03	8.90E-03	0
Dataset4	1.90E-03	4.54E+00	1.26E+00	**1.29E+270**	0	**3.43E+75**	**3.85E+06**	**3.85E+06**	0
Dataset5	7.06E-02	2.29E+00	1.01E+00	0	2.18E-02	4.00E-04	2.88E-02	2.88E-02	0
Dataset6	**1.04E+06**	2.23E+00	6.34E-01	**1.08E+160**	0	0	0	0	**Infinity**
Dataset7	9.45E+01	2.22E+00	9.04E-01	**1.93E+17**	0	0	0	0	**Infinity**
Dataset8	6.92E-01	2.55E+00	9.95E-01	**7.62E+06**	4.20E-03	9.79E+01	2.29E+00	2.29E+00	**Infinity**
Dataset9	3.40E-03	1.30E+00	8.62E-01	7.18E+01	9.46E+01	**1.09E+09**	7.70E+01	7.70E+01	**Infinity**
Dataset10	2.62E+00	9.92E-01	7.59E-01	**4.31E+39**	0	0	5.46E+00	5.46E+00	**Infinity**

2.3.3 Performance Evaluation and Discussion

I trained and built models using each dataset. The training results are shown in Table 2.3. The numbers in the parentheses are the number of correctly classified genes. As shown in the table, the accuracy of the models was at least 94%. From the training results, I concluded that the models were credible. Thus, I continued on to the next step to carry out the performance evaluation. The

evaluation and comparison results are shown in Table 2.4 and Figure 2.5.

Performance of the tools can be ranked by counting how many ground truth genes are detected by each of the DEG method. In terms of the rankings of the methods, MLDEG showed the most stable results, achieving the first place for eight datasets and the second place for two datasets. The average ranking of my method was 1.2. However, the results of the compared DEG tools varied depending on the datasets used. edgeR showed the second best performance. It ranked first for five datasets but second for three datasets and third for two datasets. The average ranking was 1.7. The average rankings of the rest of the methods were 3.4 and the rankings ranged from one to five. Admittedly, there were some datasets that all methods showed the same and best performance on such as dataset7 and dataset8 but there was no propensity observed that all of the methods worked well on some datasets and poorly on other datasets; the datasets that the methods showed good performances were all different. Meanwhile, my method outperformed the compared methods without showing such a tendency.

Nextly, as shown in Table 2.4, my method could identify almost all GT genes of six datasets: dataset1, dataset3, dataset4, dataset7, dataset8 and dataset10. The best results of each dataset are highlighted in boldface. The numbers in the parentheses are the rankings of the methods. In addition, my method won the first place in identifying GT genes of most datasets except dataset3 and dataset5; my method was next to the best for these two datasets. Meanwhile, my method showed poor performance on dataset6. However, the other four tools also showed poor performances on the same dataset. Thus, I took a careful look in the training data and the test data of dataset6. There were only one GT gene and six GT genes in the training data and the test data, respectively. One of the goals of my method is to classify the genes that are hard to be determined

whether the genes are DEGs or not. This is the baseline of the setting concept of the test data. I tried to include such genes in the test data based on the criterion that test data are the genes that are called as DEGs by at least one DEG tool. Following this criterion, my method cannot detect GT genes if the genes are not included in the test data. This aspect is directly connected to the potential weak point of my method. My method uses differential expression features derived by the DEG tools used to calculate fold changes and p-values of gene expression. Thus, the results of my method tend to be dependent on the results of the DEG tools. I leave this for my future work to improve my method to calculate differential expression features by itself.

I expected that the number of GT genes included in the training data was much more than the number of GT genes included in the test data. Interestingly, when I examined the data, it showed a tendency that GT genes were included more in the test data. For example, there were seven GT genes and 18 GT genes in the training data and test data of dataset2, respectively. There were 13 GT genes and 70 GT genes in the training data and test data of dataset9, respectively. It is encouraging because, together with the evaluation results, it becomes an important evidence that my method has the power to classify gray genes and the information contained in the network-based features shows the fact that genes work in coordinated and cooperative fashion and affect each other so considering the network information is an important task in detecting DEGs.

2.4 Conclusion

I introduce MLDEG, a machine learning-based differentially expressed gene detection method using network-based features. Due to the plethora of DEG

detection tools, users experience difficulty of choosing appropriate DEG tools. Also, they often face the problems that the results of multiple DEG tools are different so are confused which results they must trust. Moreover, setting the cutoff values of fold change and p-value is also one of the issues in using DEG tools. Usually, conventional cutoff values such as 0.5 of log 2 fold change and 0.05 of p-value are used but sometimes the number of DEGs are too huge to analyze or too small and even zero. MLDEG solved these problems as follows.

1. It captures top differentially expressed genes by integrating the results of multiple DEG tools and defines the genes as positive data.

2. It calculates network properties of the genes and carries out network propagation to measure the influence power of the genes on a network.

3. Using the integrated results in the first step and the network-based information in the second step as features, it trains a model and classifies the genes that are hard to be determined as DEGs.

Using 10 high-throughput RNA-seq datasets downloaded from GEO, I evaluated the performance of my method and compared to other DEG tools. I set GT genes by searching the original papers of the datasets and evaluated how many GT genes MLDEG could detect. My method won the first place in detecting DEGs of eight datasets and could find almost all GT genes of six datasets. From these results, I concluded that my method is powerful in identifying DEGs with robustness and competitiveness. Because of the design principle, my method can accommodate any new DEG methods naturally. I believe that biologists can use my tool without worrying about how to calibrate DEG detection methods.

Chapter 3

Construction of condition-specific pathway interaction network by computing shortest paths on weighted PPI

Identifying perturbed pathways in a given condition is crucial in understanding biological phenomena. In addition to identifying perturbed pathways individually, pathway analysis should consider interactions among pathways. Currently available pathway interaction prediction methods are based on the existence of overlapping genes between pathways, protein-protein interaction (PPI) or functional similarities. However, these approaches just consider the pathways as a set of genes, thus they do not take account of topological features. In addition, most of the existing approaches do not handle the explicit gene expression quantity information that is routinely measured by RNA-sequencing.

To overcome these technical issues, I developed a new pathway interaction network construction method using PPI, closeness centrality and shortest paths.

I tested my approach on three different high-throughput RNA-seq data sets: pregnant mice data to reveal the role of serotonin on beta cell mass, bone-metastatic breast cancer data and autoimmune thyroiditis data to study the role of IFN-α. My approach successfully identified the pathways reported in the original papers. For the pathways that are not directly mentioned in the original papers, I was able to find evidences of pathway interactions by the literature search. My method outperformed two existing approaches, overlapping gene-based approach (OGB) and protein-protein interaction-based approach (PB), in experiments with the three data sets. My results show that PINTnet successfully identified condition-specific perturbed pathways and the interactions between the pathways. I believe that my method will be very useful in characterizing biological mechanisms at the pathway level. PINTnet is available at http://biohealth.snu.ac.kr/software/PINTnet/.

3.1 Background of pathway interaction network construction

3.1.1 The importance of finding perturbed interaction between pathways

Identifying perturbed pathways in a given condition is crucial in understanding biological phenomena. Over-representation analysis (ORA) (Rivals *et al.*, 2007), gene set enrichment analysis (GSEA) (Subramanian *et al.*, 2005; Medina *et al.*, 2009; Nam *et al.*, 2010), signaling pathway impact analysis (SPIA) (Tarca *et al.*, 2009) and EnrichNet (Glaab *et al.*, 2012) are widely used approaches to identify such pathways. These approaches detect activated pathways and rank the pathways in terms of their own activation scores or statistical tests. However, pathways usually function in a coordinated and cooperative fashion (Ashwell

et al., 1996; Jamieson and Yamamoto, 2000; Itasaki and Hoppler, 2010). Thus understanding interactions or crosstalk between pathways becomes as important as identifying perturbed single pathway.

3.1.2 Challenges in pathway interaction network construction

The currently available methods are based mainly on testing differential gene expression. None of these methods use the explicit quantity of gene expression. Therefore, the methods are not able to identify the subtle but important changes in gene expression. Moreover, many of the methods do not take into account the topological features, treating the pathways just as a set of genes. Recently, a path-based approach was studied (Tegge *et al.*, 2016) but it does not use transcriptome data to predict condition-specific interactions, limiting itself to finding merely static interactions between pathways.

Here, I propose a new pathway interaction network construction method (PINTnet). The summary of my method are:

1. The interactions between pathways are represented by the subnetworks that are constructed considering two topological features: closeness centrality and shortest path.

2. Shortest paths on the subnetworks are computed based on an assumption that pathway interactions occur by a series of spontaneous reactions among genes belonging to the pathways.

3. The explicit quantity of gene expression is used to measure the activation status of pathway interactions.

4. The flow of the changes in expression is weighted. Higher weight is given to any edge between genes when the edge connects differentially expressed genes (DEGs).

3.2 Methods

In this section, I describe the process of how PINTnet measures the activation status of the interactions and constructs the pathway interaction network including the preprocess steps in detail. The overview of the method is depicted in Figure 3.1.

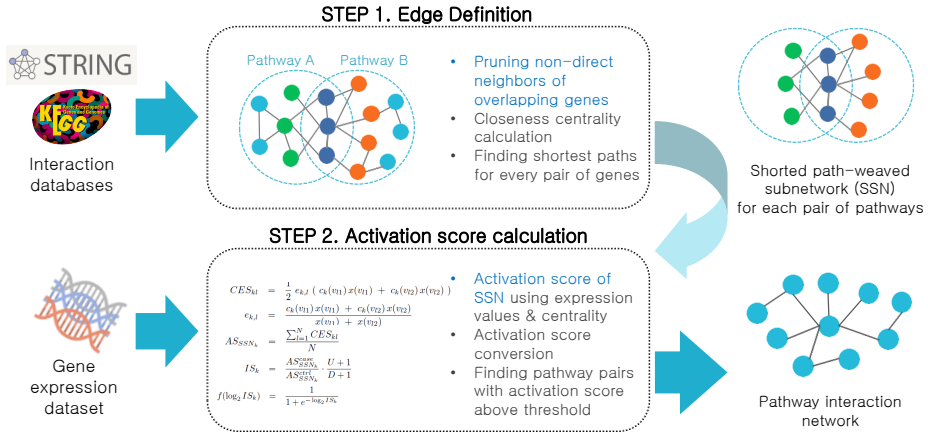


Figure 3.1: Overview of my method

3.2.1 Preparation of PPI and pathway information

I collected protein-protein interaction data from STRING (ver.9.1) (Franceschini *et al.*, 2013) and pathway data from KEGG (Release 73.1) (Kanehisa and Goto, 2000). To integrate these two independent information, I selected the genes in both datasets and edges in the pathways are augmented by bringing in edges in STRING. There are several main categories of pathways in KEGG. These are metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems and human diseases. I excluded pathways in metabolism category because the metabolic pathways focus on the metabolic products of cells and are not well represented in PPI (Hsu

et al., 2008).

3.2.2 Defining edges in the pathway network

Defining edges between two pathways is the key issue in constructing a pathway interaction network. Below are the steps for defining edges.

Step 1: Subnetwork construction on each pathway pair

I constructed a subnetwork for every possible pair of pathways. To do so, I used two criteria for the pathways to be paired: whether the two pathways have at least one overlapping gene and whether the two pathways have at least one gene connected to the overlapping genes via PPI. I defined every pair of two pathways as a possible pair only when the two pathways satisfied the both criteria. Then, for every possible pair, a subnetwork was constructed using PPI involving the two pathways.

Step 2: Closeness centrality calculation

For each subnetwork generated above, I calculated closeness centrality of all the genes within. The centrality was to evaluate the degree of a node to be central in a given network, by taking a reciprocal of an average shortest path length to all the nodes within a network from the source. The shorter the average shortest path length of a node, the closer to 1 the closeness centrality of the corresponding node is, otherwise, closer to zero. In this way, genes reflect the topological importance of themselves concerning all possible neighbor nodes within a given subnetwork.

Step 3: Shortest path computation

After calculating the closeness centrality, I pruned the genes that are not direct neighbors to overlapping genes in subnetworks. Then, I computed the shortest paths. Given two pathways A and B, let the genes in A as $A_{genes} = \{a_1, a_2, \ldots, a_m\}$ and the genes in B as $B_{genes} = \{b_1, b_2, \ldots, b_n\}$ where m is the number of genes in A and n is the number of genes in B. The shortest paths were computed for every pair of genes a_i and b_j where $1 \leq i \leq m$, $1 \leq j \leq n$ and both a_i and b_j are the direct neighbor genes to the overlapping genes. The shortest paths must pass through any overlapping gene of the two pathways.

Step 4: Constructing shortest path-weaved subnetwork

Finally, I weaved the shortest paths and constructed shortest path-weaved subnetworks (SSN). I conjectured that the pathway interaction occurs by the rapid and spontaneous flow of biological signal or interaction through topologically important genes. This concept is realized in my method by computing shortest path in the weighted subnetworks in terms of closeness centrality. The SSN thus is the network that connects the topologically important genes using the shortest paths. The overview of these steps is depicted in Figure 3.2.

Measuring activation status of pathway interaction

Measuring the activation status of biological systems or networks is technically difficult. For example, I may want to compute the average expression level of all genes in a network as the activation status of the network. However, this computation completely ignores topological features. A recent work (**?**) demonstrated that the identification and measurement of subsystems by using both PPI and pathway information were effective in prognosis of breast cancer sur-

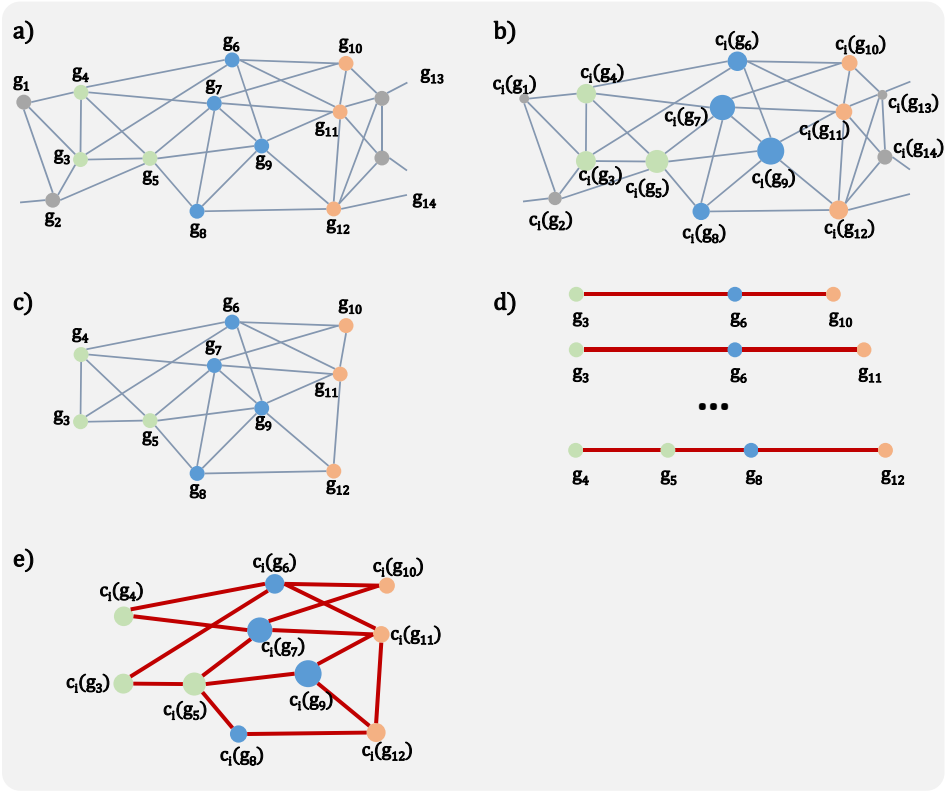


Figure 3.2: Constructing a shortest path-weaved subnetwork

g indicates genes. c_i indicates the closeness centrality of a gene of subnetwork i. Overlapping genes are colored in blue, the direct neighbors of the overlapping genes belonging to pathway A are colored in green and the direct neighbor genes belonging to pathway B are colored in orange. The others are colored in gray.

a) A subnetwork of pathway A and pathway B

b) Closeness centrality is calculated for every gene in the subnetwork. The node size represents the closeness centrality of the node.

c) The genes that are not direct neighbors to overlapping genes are pruned.

d) Shortest paths are computed.

e) The shortest paths are weaved to construct a shortest path-weaved subnetwork

vival by defining activation status of network edges. I incorporated the approach to calculate the activation status of interaction between pathways. To measure the activation status of each SSN, PINTnet firstly calculates a co-expression score (CES) of each edge of the SSN using the following equations:

$$CES_{kl} = \frac{1}{2}\, e_{k,l}\, (\, c_k(v_{l1})\, x(v_{l1}) \;+\; c_k(v_{l2})\, x(v_{l2})\,) \tag{3.1}$$

$$e_{k,l} = \frac{c_k(v_{l1})\, x(v_{l1}) \;+\; c_k(v_{l2})\, x(v_{l2})}{x(v_{l1}) \;+\; x(v_{l2})} \tag{3.2}$$

where k is the index of SSN constructed from each pathway pair, l is the index of an edge, v_{l1} and v_{l2} are two genes connected by the edge l, $c_k(v)$ is the closeness centrality of a gene v in SSN_k and $x(v)$ is the expression level of a gene v. $e_{k,l}$ is the condition-specific edge centrality of edge l in SSN_k. After measuring the co-expression score for every edge in SSN_k, PINTnet takes the average of the summation of the scores as the activation score (AS) and that is:

$$AS_{SSN_k} = \frac{\sum_{l=1}^{N} CES_{kl}}{N} \tag{3.3}$$

where N is the total number of edges in SSN_k and l is the index of an edge. PINTnet then calculates the ratio of AS_{SSN_k} for the case and the control data so it can reflect the activity of the pathway interaction in a comparative manner between case and control.

Computing DEGs is a simple but effective approach for detecting perturbed pathways and even signaling impacts in the pathways in a given condition. However, DEGs are widely interspersed and are not connected in the networks or pathways. To utilize DEG information, I applied the ratio of DEGs that are connected by edges as a weight. A higher interaction score is assigned for more DEG connections. In this step, PINTnet simply calculates the fold change of expression level of each gene to define DEGs and the default threshold is $\log_2 0.5$

as used in other studies that used RNA-seq data (Marioni *et al.*, 2008; Rapaport *et al.*, 2013; Sheikh *et al.*, 2015). The equation is as follows:

$$IS_k = \frac{AS^{case}_{SSN_k}}{AS^{ctrl}_{SSN_k}} \cdot \frac{U+1}{D+1} \tag{3.4}$$

where k is the index of SSN, IS_k is the interaction score of SSN_k, $AS^{case}_{SSN_k}$ is the activation score of SSN_k of the case data, $AS^{ctrl}_{SSN_k}$ is the activation score of SSN_k of the control data, U is the number of connected up-regulated DEGs of the case data compared to the control data and D is the number of connected down-regulated DEGs of the case data compared to the control data. When PINTnet calculates the fold change, the cutoff value of 1.0 for the expression level is used to prevent noise such as extremely high fold change due to the comparison between small numbers. The cutoff value was set based on other studies (Brooks *et al.*, 2011; Shin *et al.*, 2014). In addition, genes that are overlapped among multiple pathways can cause false positives. A study reported this issue and proposed an approach of ruling out the overlapping genes when determining perturbed pathways (Donato *et al.*, 2013). I tried to attenuate the effect of those genes by dividing the expression level by the number of pathways that the genes belong to, so that it could be naturally considered in calculating the ratio of connected DEGs.

Pathway interaction network construction

After measuring the activation status of all pairs of pathways and obtaining the interaction score, PINTnet converts the interaction score using the sigmoid function (Lever *et al.*, 2016). It is to convert the scores to a value in the range between 0 and 1, so a constant cutoff value can be applied uniformly to all SSNs to construct the pathway interaction network using the only pairs satisfying the cutoff. The input value of sigmoid function must be between -1 and 1 so

PINTnet takes log of 2 of the interaction scores. The equation of the function is as follows:

$$f(\log_2 IS_k) = \frac{1}{1 + e^{-\log_2 IS_k}} \tag{3.5}$$

After the interaction score is converted, a pathway interaction network is constructed with the edges between pathways when the interaction score of edges satisfies the cutoff value. I empirically determined the cutoff value by testing PINTnet on various data from other biological researches.

3.3 Results

To evaluate the performance of PINTnet, I used three different high-throughput RNA-seq datasets in Gene Expression Omnibus (GEO). Cytoscape was used to visualize the networks (Shannon *et al.*, 2003). The test datasets are summarized in Table 3.1. For the evaluation, I investigated the evidences for every edge that connected the pathways reported in the original papers through the literature search and established the evaluation criteria for the performance of PINTnet and two existing pathway interaction network construction methods, overlapping gene-based approach (OGB) and PPI-based approach (PB), were used for the performance comparison. The details of the approaches are described in Performance comparison to other methods section.

3.3.1 Data description

Dataset1 is the data that measured the gene expression levels of pregnant mice to reveal how serotonin regulates pancreatic beta cell mass during pregnancy (Kim *et al.*, 2010). The authors compared the global gene expression patterns in islets from nonpregnant and pregnant female mice by the high-throughput sequencing to identify the genes potentially involved in regulating maternal beta

Table 3.1: The description of three datasets

Name	Title	Accession No.
Dataset1	Serotonin regulates pancreatic beta cell mass during pregnancy	GSE21860
Dataset2	ABL kinases promote breast cancer osteolytic metastasis	GSE69125
Dataset3	IFN-α mediates the development of autoimmunity	GSE25115

cell mass. They stated that *Tph1* and *Tph2* were the genes most markedly induced during pregnancy. These two genes encode two isoforms of tryptophan hydroxylase, the rate-limiting enzyme in the synthesis of serotonin, 5-HT. The authors also reported that beta cells share a common gene expression program and the ability to synthesize, store and secrete serotonin with serotonergic neurons.

Dataset2 is the data generated by a study investigating how ABL kinases promote breast cancer osteolytic metastasis (Wang *et al.*, 2016b). Bone is one of the primary sites where breast cancer metastasizes and 70% of deaths of breast cancer is caused by bone metastases. The authors evaluated the result of single- or double-knockdown of ABL1 and ABL2 in breast cancer cells using RNA-seq analysis to reveal the signaling pathways required for ABL kinases-dependent bone metastasis. They carried out GSEA to identify which pathways were affected by ABL kinases in metastatic breast cancer cells. They reported that Jak-STAT signaling pathway, Hippo signaling pathway, cytokine-cytokine interaction and bone metastasis were enriched in the control group compared to ABL1/ABL2 knockdown group.

Dataset3 is from a study that used thyroiditis as a model to reveal how IFN-α plays a pivotal role in auto immunity (Akeno *et al.*, 2011). The authors generated transgenic mice overexpressing IFN-*alpha* in the thyroid and performed RNA-seq analysis. The transgenic mice showed upregulation of pathways such as antigen presentation pathway, interferon signaling, complement system, apoptosis, pattern recognition receptors and RAR activation.

3.3.2 Evaluation criteria

Dataset1

It is well known that nutrient requirements by the fetus incur change in the maternal metabolism during pregnancy. Nutrient flow to the fetus is maintained by increasing insulin resistance. The resistance may cause maternal hyperglycemia but the glucose level is maintained by the expansion of beta cells driven by prolactin and placental lactogen (Assche *et al.*, 1978; Parsons *et al.*, 1992; Huang *et al.*, 2009). Failures in this response raises the risk of being diagnosed with gestational diabetes mellitus (Rieck and Kaestner, 2010). Serotonin is a regulator of insulin secretion and co-localized with insulin in granules of pancreatic β-cells. A lack of serotonin in β-cells can lead to reduced insulin secretion (Paulmann *et al.*, 2009). Also, it is known that prolactin has direct effects on increasing insulin secretion (Nielsen, 1982; Sorenson *et al.*, 1987; Bole-Feysot *et al.*, 1998) and is closely related to diabetes (Bernard *et al.*, 2015).

Dataset2

Ras signaling pathway is the pathway which *ABL1* and *ABL2* belong to and it is known that Ras signaling pathway activation is implicated in breast cancer invasion and growth (Karnoub and Weinberg, 2008). Thus the downstream of

Ras signaling is considered to be a potential target against osteolytic breast cancer metastasis (Bosma *et al.*, 2014). MAPK signaling pathway is known to be implicated in cancer-induced bone pain (Sukhtankar *et al.*, 2011). In addition, it is known that p38 MAPK is important in maturation and synthesis of osteoclasts (Matsumoto *et al.*, 2000; Zwerina *et al.*, 2006). Wnt signaling pathway is one of the pathways dysregulated in human breast cancer and it was reported that the activity of Wnt signaling in breast cancer is significantly higher than that in bulk cancer cells (Jang *et al.*, 2015). Upregulation of Wnt signaling pathway has been reported to lead to increased metastasis including bone metastasis from breast cancer (Chen *et al.*, 2011; Dey *et al.*, 2013). TGF-β signaling pathway was reported to be important for the development of osteolytic bone metastases by numerous studies (Buijs *et al.*, 2012). Proteoglycans participate in the control of bone tumor development and bone metastases dissemination (Velasco *et al.*, 2010). A high sensitivity to PI3K-Akt signaling pathway characterizes triple-negative breast cancer metastasis to bone (Guise, 2013). Hippo signaling pathway deregulation in breast cancer bone metastasis has been suggested that YAP and TAZ activity was increased in metastatic breast cancer (Cordenonsi *et al.*, 2011). In addition to the individual functions of the pathways, interactions between the pathways are reported by various studies (Deel *et al.*, 2015; Pan, 2010; Aksamitiene *et al.*, 2012; Guardavaccaro and Clevers, 2012; Vadlakonda *et al.*, 2014; Guo and Wang, 2009; Rawlings *et al.*, 2004; Pataki *et al.*, 2015; Iozzo and Sanderson, 2011).

Dataset3

It is known that Toll-like receptor signaling pathway plays an important role in autoimmunity including thyroid autoimmunity (Toubi and Shoenfeld, 2004; Kawashima *et al.*, 2011). Also, antigen presentation, complement system, apop-

tosis and pattern recognition receptors are known to involve in thyroid autoimmunity (Feldmann *et al.*, 1992; Potlukova and Limanova, 2006; Wang and Baker Jr, 2007; Merrill and Mu, 2015).

3.3.3 Performance comparison to other methods

To compare the performance of PINTnet to other approaches, I implemented the overlapping gene-based approach (OGB) and the PPI-based approach (PB) and ran three methods including mine on the three test datasets in the previous sections.

Overlapping gene-based approach

This method is a two-step approach. In the first step, the activation status of each pathway was calculated using Fisher's exact test with a contingency table dealing with two parameters: one is whether a gene is a DEG or not and the other is whether a gene belongs to the pathway or not. Then, in the second step, the significance of edges among pathways was evaluated using Fisher's exact test. The significance of pathways and the edges among the pathways were determined at a p-value of 0.05 or less. The significant edges were used to construct a pathway network.

PPI-based approach

For PPI-based approach, I implemented the simple version of the approach since no executable code is available. I implemented the approach based on the hypothesis that the more interactions may guarantee the higher probability of interaction. To do this, I calculated the empirical p-values for every possible pair of pathways to find significantly interacting pathway pairs by shuffling the original PPI network 1,000 times, counting the number of round when the

number of shuffled PPI edges between pathways was bigger than or equal to that of the original PPI edges and dividing the number by the number of round, in this case, 1,000. Then I adjusted the p-values using Bonferroni correction and took the edges with the p-value less than 0.05 as the significantly interacting edges. Connecting the edges, I constructed a template network and calculated active PPIs on the network using the datasets.

Running the approach on the test datasets, I observed that too many nodes and edges were connected even though multiple testing correction was performed using the Bonferroni correction. For example, there were 271 nodes and 12,264 edges for *dataset1*. It seemed almost impossible to determine which pathways and interactions between the pathways were important in the given conditions. Thus I did not include this approach for the performance comparison.

Comparison results

I compared the performance of the approaches based on the biological evidences found by the literature search and organized in Evaluation criteria section. The following criteria were used for the quantitative measure of the performance. The first criterion was the interactions between the pathways. I calculated the percentage of the number of the evidence-supported edges between the evidence-supported pathways against the total number of edges in the network. It was to measure how successfully the approaches connected the correct edges supported by evidence. The second criterion was the degree of pathways. I calculated the ratio of the average degree of the evidence-supported pathways against that of all pathways in the network. It was to measure how the approaches placed the important pathways as hubs in the central position of the network. The last criterion was to see how successfully the approaches rescued the correct pathways.

I calculated the percentage that how many evidence-supported pathways were rescued. I confirmed that PINTnet surpasses other methods from the results described in the following paragraphs and the comparison results are shown in Table 3.2 and Figure 3.3.

In the pathway interaction network constructed using my method on *dataset1* with the cutoff value of 0.95, there were 60 pathways and 92 edges between the pathways. The network is shown in Figure 3.4. Among the pathways, serotonergic synapse (mmu04726), insulin secretion (mmu04911), insulin resistance (mmu04931), prolactin signaling pathway (mmu04917), pancreatic secretion (mmu04972) and three diabetic pathways (mmu04930, mmu04940 and mmu04950) were included as nodes. In addition, it was observed that the edges in the network connected insulin secretion and serotonergic synapse, insulin secretion and pancreatic secretion, insulin secretion and prolactin signaling pathway, insulin resistance and diabetes-related pathways, and prolactin signaling pathway and diabetes-related pathways. The interactions between the pathways suggest how biological response occur during pregnancy by the cooperative work of relevant pathways. In addition, these interactions may give a point of view to conceive how the interactions of pathways drive the expansion of beta cell mass. The edges connected to diabetic pathways may imply the high chances of being diagnosed with gestational diabetes mellitus due to insulin resistance. Meanwhile, OGB failed to detect insulin resistance and one of the diabetic pathways even though there were 93 pathways and 109 edges. Also, there was only one edge connecting two remaining diabetic pathways.

There were 66 pathways and 122 edges between the pathways in the pathway interaction network constructed using my method on *dataset2* with the cutoff value of 0.99. The network is shown in Figure 3.5. I observed Jak-STAT signaling pathway (hsa04630), Hippo signaling pathway (hsa04390), cytokine-cytokine re-

Table 3.2: Comparison results

(a) The number of edges between the pathways in the pathway interaction network. The first column of each approach is the number of edges between the evidence-supported pathways. The second column of each approach is the number of edges between all pathways in the network.

(b) The average degree of the pathways in the pathway interaction network. The first column of each approach is the average degree of the evidence-supported pathways. The second column of each approach is that of all pathways in the network.

(c) The number of evidence-supported pathway found in the network. The first column is the number of evidence-supported pathways that are found in the pathway interaction network constructed using my method. The second column is the number of evidence-supported pathways that are found in the pathway interaction network constructed using OGB. The third column is the number of all evidence-supported pathways.

(a)

Data	PINTnet		OGB	
	Evidence-supported	All	Evidence-supported	All
Dataset1	9	92	1	109
Dataset2	15	122	2	268
Dataset3	1	149	1	291

(b)

Data	PINTnet		OGB	
	Evidence-supported	All	Evidence-supported	All
Dataset1	9.500	3.067	3.800	2.344
Dataset2	9.700	3.697	4.000	3.829
Dataset3	5.000	2.922	6.000	4.376

(c)

Data	Found		All
	PINTnet	OGB	
Dataset1	8	6	8
Dataset2	10	9	10
Dataset3	6	4	6

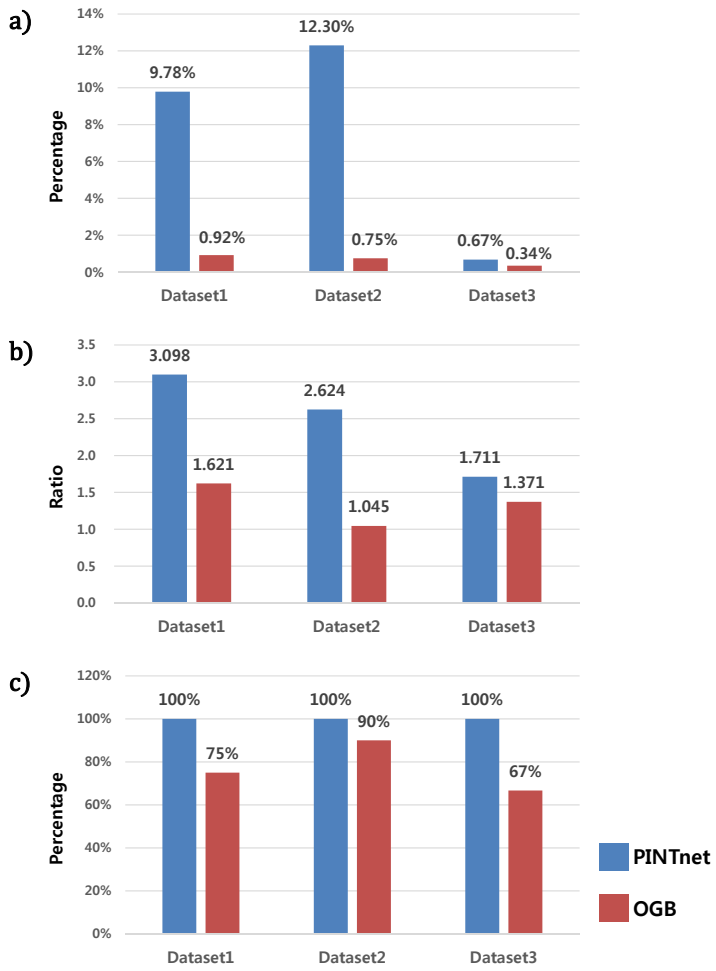


Figure 3.3: Comparison results

(a) is the percentage of the number of evidence-supported edges against the number of all edges in the pathway interaction network. PINTnet outperformed OGB in identifying the edges connected by the evidence-supported pathways.

(b) is the ratio of the average degree of the evidence-supported pathways and that of all pathways in the pathway interaction network. The evidence-supported pathways had more edges when detected by PINTnet than detected by OGB.

(c) is the percentage of how many evidence-supported pathways are found in the pathway interaction network.

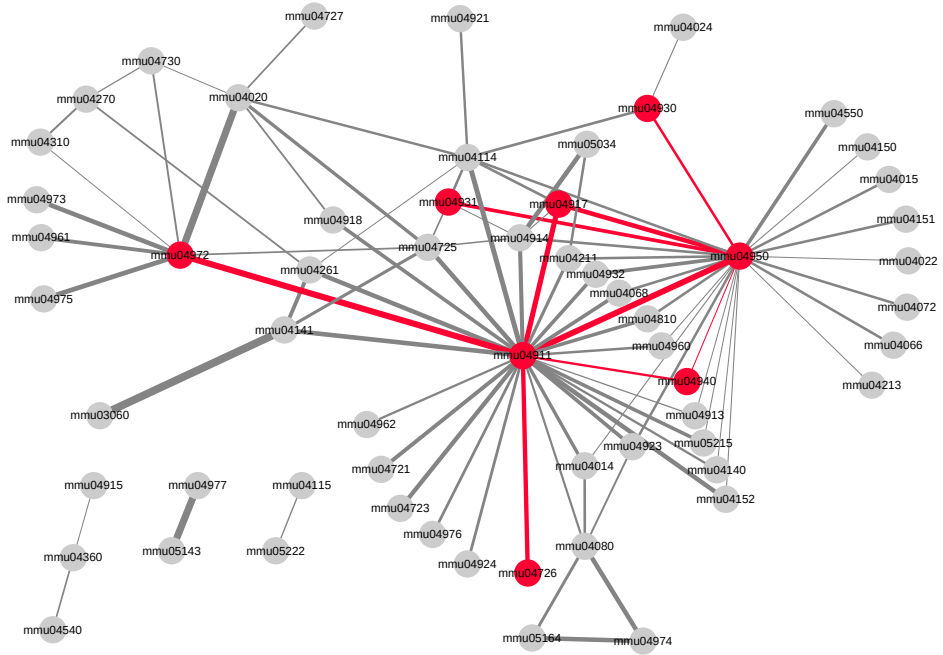


Figure 3.4: A pathway interaction network of pregnant mice

60 pathways are connected by 92 edges in this network. The pathways that coincide with the result of the original paper are rescued. The pathways are serotonergic synapse (mmu04726), insulin secretion (mmu04911), prolactin signaling pathway (mmu04917), pancreatic secretion (mmu04972), insulin resistance (mmu04931) and three diabetic pathways (mmu04930, mmu04940 and mmu04950) and colored in red. The edges connecting these pathways are also colored in red. The width of edges is set according to the activation score. The higher the activation score, the thicker the edge.

ceptor interaction (hsa04060) and osteoclast differentiation (hsa04380), which were reported by the original paper, were included in the network. In addition to the pathways, I identified the pathways that I found the evidences of functional importance in bone-metastatic breast cancer from the literatures lying on the multiple paths from Ras signaling pathway (hsa04014) to osteoclast differentiation. The pathways were MAPK signaling pathway (hsa04010), Wnt signaling pathway (hsa04310), Hippo signaling pathway, TGF-β signaling pathway (hsa04350), PI3K-Akt signaling pathway (hsa04151) and proteoglycans in cancer (hsa05205). The result implies that the pathways implicated in bone metastasis from breast cancer interact each other and the interactions among the pathways along with the paths may give the insight of how bone metastatic breast cancer is caused by pathways interaction. However, TGF-β signaling pathways was not rescued by OGB and only two edges were detected: MAPK signaling pathway and proteoglycans in cancer; Ras signaling pathway and PI3K-Akt signaling pathway.

The pathway interaction network constructed using my method on *dataset3* with the cutoff value 0f 0.99 included 102 pathways and 149 edges between the pathways. The network is shown in Figure 3.6. The network successfully included the pathways that mentioned to be upregulated in the original paper except RAR activation because there is no proper match of the pathway in KEGG. The pathways were Toll-like receptor signaling pathway (hsa04620), autoimmune thyroid disease (hsa05320), complement and coagulation cascades (hsa04610), antigen processing and presentation (hsa04612), apoptosis (hsa04210) and RIG-I-like receptor signaling pathway (hsa04622). There was only one edge between the pathways and Toll-like receptor signaling pathway and autoimmune thyroid disease were connected by the edge. However, Toll-like receptor signaling pathway is the pathway that IFN-α belongs to and autoimmune thyroid

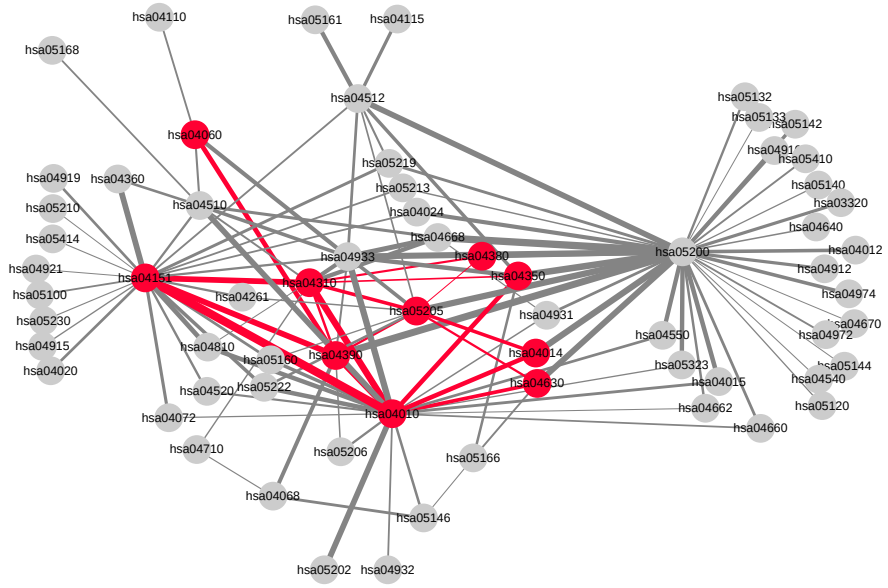


Figure 3.5: A pathway interaction network of bone metastasis from breast cancer

66 pathways are connected by 122 edges in this network. The original paper reported Jak-STAT signaling pathway (hsa04630), cytokine-cytokine receptor interaction (hsa04060), Hippo signaling pathway (hsa04390) and bone metastasis were upregulated in the control compared to ABL1/ABL2 knockdown mice. I found multiple paths from Ras signaling pathway (hsa04014), ABL kinases belong to, to osteoclast differentiation (hsa04380) through MAPK signaling pathway (hsa04010), Wnt signaling pathway (hsa04390), TGF-β signaling pathway (hsa04350), PI3K-Akt signaling pathway (hsa04151), Hippo signaling pathway (hsa04390) and proteoglycans in cancer (hsa05205). I found the evidences in literature that these pathways are related to bone metastasis from breast cancer. These pathways and the edges between the pathways are colored in red and the width of edges are set according to the activation score.

disease is the overall context of the original paper. Moreover, the findings reported in (Toubi and Shoenfeld, 2004; Kawashima *et al.*, 2011) supported that Toll-like receptor signaling pathway plays an important role in autoimmunity as mentioned in Evaluation criteria section. On the contrary, OGB failed to detect RIG-I-like receptor signaling pathway and complement and coagulation cascades. In addition, there was only one edge between antigen processing and presentation and autoimmune thyroid disease.

3.4 Discussion

Currently available approaches for constructing pathway network were designed to handle microarray data so the approaches mostly rely on the statistical tests. The approaches determine the significance of the interaction by the p-value yielded by the tests or use the p-value itself to calculate the secondary score for the determination. In addition, even though several approaches incorporated PPI to infer the interactions between pathways, the approaches have a limitation that PPI was treated merely as a set of individually represented genes without considering any relation between the genes. To address these issues, I applied the concept of closeness centrality and shortest paths to define the edges in the pathway interaction network. I assumed that the interaction between two pathways will occur when the biological signals rapidly flow through the topologically important genes. Based on the assumption, I constructed shortest path-weaved subnetworks to represent the edges and calculated interaction score using explicit gene expression quantity on the subnetworks.

The scoring scheme of PINTnet is a ranking method. It constructs the pathway interaction network using pairs of pathways of which the score is higher than the cutoff value. The results on the test datasets suggest that PINTnet

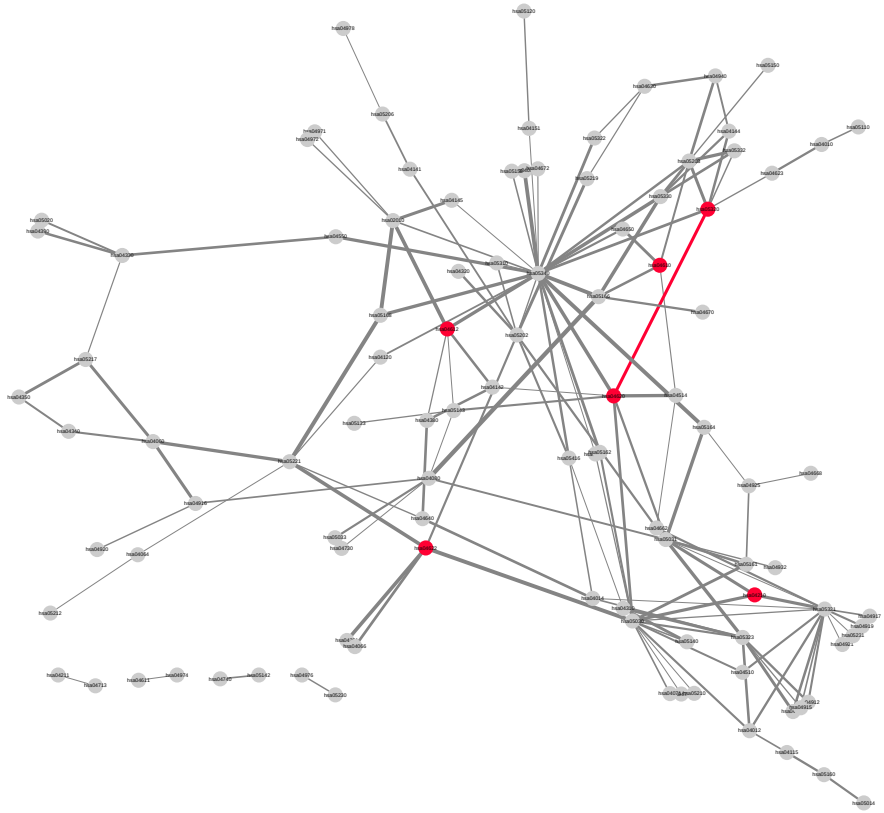


Figure 3.6: A pathway interaction network of IFN-α mediated autoimmunity

102 pathways are connected by 149 edges in this network. The original paper reported that Toll-like receptor signaling pathway (hsa04620), complement and coagulation cascades (hsa04610), antigen processing and presentation (hsa04612), RIG-I-like receptor signaling pathway (hsa04622) and apoptosis (hsa04210) were upregulated and my method rescued the pathways including autoimmune thyroid disease (hsa05320). There is only one edge connecting these pathways and the edge connects Toll-like receptor signaling pathway and autoimmune thyroid disease.

successfully reproduced the results of the original papers and, therefore, is useful in analyzing the perturbed pathways and their interactions in a given condition.

Like existing methods, PINTnet is based on the identification of overlapping genes between two pathways. I assumed that the overlapping genes function as a bridge between two pathways. Based on the assumption, I considered the situation that at least one overlapping gene exists as one of the rules to define the edge in the pathway interaction network. This criterion, though reasonable and popular, may be too stringent. For example, when two pathways are well connected by direct edges in PPI but do not share any genes, it is not clear whether the two pathways interact or not. Therefore, the pairs of truly interacting pathways might be ruled out. I will further work on the matter to overcome the limitation.

3.5 Conclusion

In this work, I developed a new pathway interaction network construction method, PINTnet. Running PINTnet on the three datasets to test the performance, I observed that it successfully rescued the findings reported in the original papers. In the result of *dataset1*, PINTnet successfully detected the pathways related to the changes occurring during pregnancy. Also I observed that the pathways were connected by the edges supported by the literatures. For *dataset2*, I also identified that the pathways related to bone-metastatic breast cancer were rescued in the pathway interaction network and the edges between the pathways implied the interactions participating in the induction of the phenotype. For *dataset3*, the pathways reported by the original paper were included as nodes in the pathway interaction network and there was a connected edge between Toll-like receptor signaling pathway and autoimmune thyroid disease.

I expect PINTnet to be a useful tool for pathway interaction network analysis. PINTnet is available at http://biohealth.snu.ac.kr/software/PINTnet/.

Chapter 4

Bioinformatics analyses with peripheral blood RNA-sequencing unveiled the cause of the graft loss after pig-to-nonhuman primate islet xenotransplantation model

Clinical islet transplantation is a promising treatment option for intractable type 1 diabetes. Although short term result of islet function after transplantation has been improved, outcome of long-term islet graft function is still unsatisfactory. The causes of this islet graft loss in the chronic phase are obscure. In pre-clinical islet xenotransplantation, since consistent long-term islet graft survival had not been achieved yet, it was impossible to explore the mechanism of chronic phase islet graft loss so far. However, recent consistent long-term survivals of adult porcine islets $\geq$ 6 months in five independent diabetic nonhuman primates (NHPs) enabled me to investigate the cause of chronic phase islet graft loss in xenotransplantation. I sought to analyze the gene expression profiles

using peripheral blood RNA-sequencing to find out potential cause(s) chronic graft failure. Bioinformatics analyses showed that highly relevant 'immunologic' pathways were activated in NHP experienced chronic phase graft failure before the overt graft failure. Further connectivity analyses revealed that activation of T-cell signaling pathways was the most prominent, suggesting T-cell mediated graft rejection could be the cause of the chronic phase islet loss. Indeed, the porcine islets heavily infiltrated with CD3+ T cells on biopsied liver samples confirmed the T cell mediated graft rejection. Furthermore, hypothesis test with computational experiment reinforced my conclusion. Taken together, I suggested that bioinformatics analyses with peripheral RNA sequencing unveiled the cause of insidious chronic islet graft loss.

4.1 Background

Since Edmonton protocol was introduced in 2000 (Shapiro *et al.*, 2000), human pancreatic islet transplantation has become an established treatment option for type 1 diabetic patients who frequently experience fatal hypoglycemic unawareness. However, over half of the patients transplanted with human islet returned to be insulin-dependent state within 5 years (Ryan *et al.*, 2005; Barton *et al.*, 2012). The causes for this chronic islet graft loss are controversial. They encompass a higher rate of islet apoptosis due to ER stress (Fonseca *et al.*, 2011; Negi *et al.*, 2012), hypoxia (Lau *et al.*, 2009; Zheng *et al.*, 2012) in end-portal venules in the liver, and recurrent autoimmunity (Pugliese *et al.*, 2011). In addition, there is evidences that metabolic deterioration due to lipid accumulated around the islets (lipotoxicity) (Lee *et al.*, 2007; Leitão *et al.*, 2010) and toxicity of immunosuppressive drugs (Barlow *et al.*, 2013; Drachenberg *et al.*, 1999) can result in graft loss. Also, insufficient immune suppression could also be an

important cause of chronic islet loss. However, none of the above can clearly explain the exact causes of chronic islet graft loss. Recently, the researchers who participated in a research with me reported consistent long-term pig islet graft survivals ≥ 6 months in five independent monkeys. This unique opportunity allows them to examine how the pig islets are lost in the chronic phase after islet transplantation. Here, they selected two monkeys with same immunosuppressive regimen to analyze the cause of chronic islet loss in xenotransplantation; one (R051) had stable graft function for entire follow-up periods and the other (R080) lost graft function around 160 days post transplantation (DPT)

4.2 Results

4.2.1 Peripheral blood RNA sequencing

R051 showed complete normal glycemia and glucose disposal capacity for entire follow-up periods, whereas the other (R080) exhibited relatively early hyperglycemia around 160 days post transplantation (DPT), suggesting a graft failure. Intra-venous glucose tolerance test (IVGTT) showed that the pig islet graft loss in R080 was in progress between DPT120 and 180 (Figure 4.1 a)-d), processed from published data (Shin *et al.*, 2015)). Vital signs, routine laboratory examinations including complete blood cell count (CBC), liver function test (LFT), C-reactive protein (CRP), kidney function test (blood urine nitrogen/creatinine), electrolyte panel (sodium, potassium and chloride), lipase and amylase showed no abnormal findings in both of the monkey (data not shown). Also, peripheral blood lymphocyte subsets monitoring by flow cytometry, titer of donor-specific antibody by enzyme-linked immunosorbent assay (ELISA), peripheral blood cytokine level by cytometric bead assay (CBA) did not reveal any noticeable changes (data not shown). Since recent report showed gene ex-

pression perturbation in peripheral blood could reflect graft site event (Chen *et al.*, 2013; Dorr *et al.*, 2015), RNA sequencing was performed with archives of whole blood samples taken at four different time points from graft-losing R080 versus graft-stable R051 to explore the cause of chronic islet loss happened in R080 (Figure 4.1 e)).

4.2.2 Graft loss period-related activated pathways (GLPAPs) defined by TRAP (Time-series RNA-seq analysis package)

After confirming the validity of RNA-seq data, I used Time-series RNA-seq Analysis Package (TRAP) (Jo *et al.*, 2014) to determine which pathways play roles in graft loss process. Because t_2 and t_3 represent the graft-maintaining and the graft-losing period, respectively in R080, I focused on these time points and selected pathways as follows: i) select up-regulated pathways with p-value under 0.05 from the results yielded by TRAP comparing t_3 and t_2 in R080, ii) select up-regulated pathways with p-value under 0.05 from the results yielded by comparing R080 and R051 at t_3 and then take the pathways belonging to the intersection of those sets (Figure 4.2). As a result, I could obtain 59 pathways among 287 pathways in Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) Rhesus database and these pathways were dubbed graft loss period-related activated pathways (GLPAPs) as can be seen in Table 4.1.

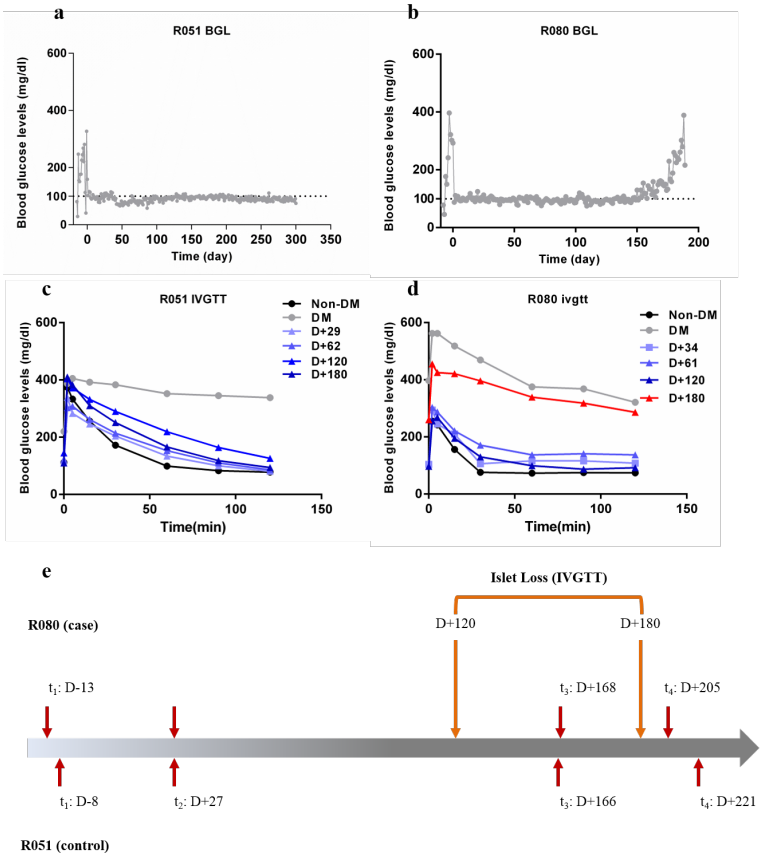


Figure 4.1: Graft function, cell-mediated immune response monitoring and experimental scheme

a) and b) indicate the glucose level of R051 and R080, respectively. R080 showed gradual increase of blood glucose level around DPT 150.

c) and d) indicate the IVGTT results of R051 and R080, respectively. Between DPT 120 and 180, R080 showed prominent glucose intolerance.

e) Sampling time points for RNA-sequencing. Whole blood archives were used for RNA sequencing. (t_1: before transplantation, t_2: one month after transplantation, t_3: immediate after increase of blood glucose in R080 and corresponding time point for R051, t_4: after overt hyperglycemia in R080 and corresponding time point for R051)

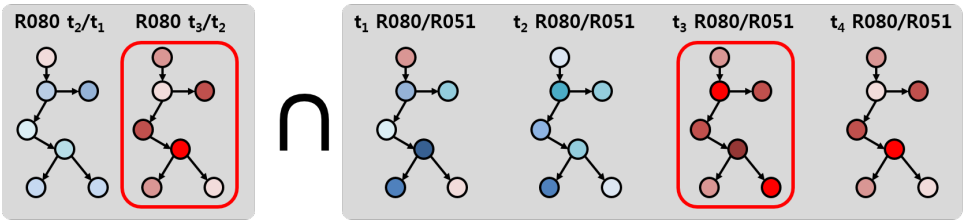


Figure 4.2: Pathway filtering strategy

GLPAPs were selected by taking the intersection of two TRAP results. One was to detect which pathways were activated at t_3 compared to t_2 in R080. The other was to detect which pathways were activated in R080 compared to R051 at t_3.

Table 4.1: Graft losing period-related activated pathways (GLPAPs)

59 out of 287 pathways in rhesus KEGG database were selected after applying of TRAP algorithm.

Pathway	Name	Category
mcc04062	Chemokine signaling pathway	Immune system
mcc04611	Platelet activation	Immune system
mcc04620	Toll-like receptor signaling pathway	Immune system
mcc04621	NOD-like receptor signaling pathway	Immune system
mcc04623	Cytosolic DNA-sensing pathway	Immune system
mcc04650	Natural killer cell mediated cytotoxicity	Immune system

Continued on next page

Table 4.1 – continued from previous page

Pathway	Name	Category
mcc04660	T cell receptor signaling pathway	Immune system
mcc04662	B cell receptor signaling pathway	Immune system
mcc04664	Fc epsilon RI signaling pathway	Immune system
mcc04670	Leukocyte transendothelial migration	Immune system
mcc04010	MAPK signaling pathway	Signal transduction
mcc04012	ErbB signaling pathway	Signal transduction
mcc04022	cGMP-PKG signaling pathway	Signal transduction
mcc04064	NF-kappa B signaling pathway	Signal transduction
mcc04068	FoxO signaling pathway	Signal transduction
mcc04070	Phosphatidylinositol signaling system	Signal transduction
mcc04152	AMPK signaling pathway	Signal transduction
mcc04370	VEGF signaling pathway	Signal transduction
mcc04630	Jak-STAT signaling pathway	Signal transduction
mcc04668	TNF signaling pathway	Signal transduction
mcc04910	Insulin signaling pathway	Endocrine system

Continued on next page

Table 4.1 – continued from previous page

Pathway	Name	Category
mcc04915	Estrogen signaling pathway	Endocrine system
mcc04917	Prolactin signaling pathway	Endocrine system
mcc04918	Thyroid hormone synthesis	Endocrine system
mcc04919	Thyroid hormone signaling pathway	Endocrine system
mcc04921	Oxytocin signaling pathway	Endocrine system
mcc03013	RNA transport	Translation
mcc04210	Apoptosis	Cell growth and death
mcc05211	Renal cell carcinoma	Cancer: Specific types
mcc05212	Pancreatic cancer	Cancer: Specific types
mcc05213	Endometrial cancer	Cancer: Specific types
mcc05214	Glioma	Cancer: Specific types
mcc05215	Prostate cancer	Cancer: Specific types
mcc05219	Bladder cancer	Cancer: Specific types
mcc05220	Chronic myeloid leukemia	Cancer: Specific types
mcc05221	Acute myeloid leukemia	Cancer: Specific types
mcc05223	Non-small cell lung cancer	Cancer: Specific types
mcc04141	Protein processing in endoplasmic reticulum	Folding, sorting and degradation

Continued on next page

Table 4.1 – continued from previous page

Pathway	Name	Category
mcc04320	Dorso-ventral axis formation	Development
mcc04380	Osteoclast differentiation	Development
mcc04540	Gap junction	Cellular communication
mcc04810	Regulation of actin cytoskeleton	Cell motility
mcc04961	Endocrine and other factor-regulated calcium reabsorption	Excretory system
mcc04722	Neurotrophin signaling pathway	Nervous system
mcc04725	Cholinergic synapse	Nervous system
mcc04060	Cytokine-cytokine receptor interaction	Signaling molecules and interaction
mcc05142	Chagas disease (American trypanosomiasis)	Infectious diseases: Parasitic
mcc05143	African trypanosomiasis	Infectious diseases: Parasitic
mcc05144	Malaria	Infectious diseases: Parasitic
mcc04623	Hepatitis B	Infectious diseases: Viral
mcc04650	Measles	Infectious diseases: Viral
mcc04660	Influenza A	Infectious diseases: Viral
mcc04662	HTLV-I infection	Infectious diseases: Viral
mcc04664	Herpes simplex infection	Infectious diseases: Viral

Continued on next page

Table 4.1 – continued from previous page

Pathway	Name	Category
mcc04670	Epstein-Barr virus infection	Infectious diseases: Viral
mcc04970	Salivary secretion	Digestive system
mcc05200	Pathways in cancer	Cancers: Overview
mcc05203	Viral carcinogenesis	Cancers: Overview
mcc05205	Proteoglycans in cancer	Cancers: Overview

After obtaining 59 of GLPAPs, p-values for each 'category' of the pathways were calculated using Fisher's exact test to determine how significantly GLPAPs were enriched in each category. To calculate p-values, I constructed a contingency table with two variables: GLPAP and category. Each cell of the table was filled by the number of the pathways according to the standard if the pathway belongs to the category or not and if the pathway is GLPAP or not (Figure 4.3). The p-values for each category are shown in Table 4.2. The most enriched category was found to be "immune system" despite any perturbation of immunological parameters in routine immune monitoring system was not found. This finding strongly implies that immunological responses are somehow activated and ongoing during t_3 in R080 compared to t_2 in R080 and corresponding time points in R051.

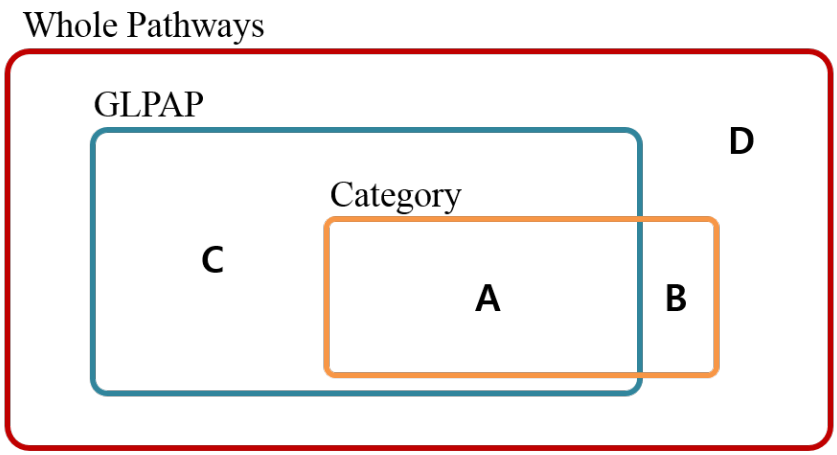
4.2.3 Pathway interaction network analysis

Even though I found that the pathways categorized in immune system were enriched mostly after GLPAPs filtering, I was not able to specify single pathways, which are potentially responsible for graft loss. Because biological pathways usually function in a cooperative manner by constituting a network, understanding

Table 4.2: Significantly enriched categories of GLPAPs

Categories are listed in ascending order of p-values calculated by Fisher's exact test. Immune system category pathways were highly enriched.

Category	**P-value**
Immune system	0.0001962
Cancers: Specific types	0.0003236
Infectious diseases: Viral	0.0003591
Signal transduction	0.0120207
Infectious diseases: Parasitic	0.1036661
Endocrine system	0.1076348
Development	0.1083940
Cell motility	0.2055749
Cancers: Overview	0.2132333
Digestive system	0.6910951
Nervous system	1.0000000
Cell growth and death	1.0000000
Cellular communication	1.0000000
Excretory system	1.0000000
Folding, sorting and degradation	1.0000000
Signaling molecules and interaction	1.0000000
Translation	1.0000000



Fisher's exact test

	GLPAP	Not GLPAP
In Category	A	B
Not in Category	C	D

Figure 4.3: A contingency table with two variables to calculate p-values of each category

To calculate GLPAP enrichment for each category of pathways, I built a contingency table for each category according to the two variables. One is whether a pathway is GLPAP or not and the other is whether a pathway belongs to the concerned category.

the network of pathways can provide the insight about which pathways are important in a given condition. Therefore, it would be desirable to analyze the network of the pathways to find out the most interacting pathways to induce graft loss among GLPAPs. To this end, I constructed a pathway interaction network of GLPAPs using PINTnet (Moon *et al.*, 2017). There were 52 pathways out of 59 GLPAPs connected by 225 edges in the network (Figure 4.4). I calculated closeness centrality for every node and used degree information to analyze which pathways played a central role in the network to induce the biological response at t_3 of R080. I focused only on the pathways belonging to immune system because immune system was the most enriched category as mentioned in the previous section. The pathways with the closeness centrality value and degree higher than the average closeness centrality value and the average degree of all the nodes in the network were considered meaningful. Among eight pathways of immune system, three met the criteria and the pathways were T cell receptor signaling pathway, B cell receptor signaling pathway, and Platelet activation. The pathways are shown in Table 4.3. The results suggested that T cell mediated immune rejection toward the pig islets was in progress at t_3 of R080. Liver biopsy samples at DPT184 from R080 from the archives were collected and graft histology was examined by immunohistochemistry. It was found that insulin positive islet grafts were heavily infiltrated by mostly CD3+ T cells (Figure 4.5).

4.2.4 Hypothesis evaluation using network propagation

To reconfirm my findings, I sought to test each hypothesis which could explain islet loss. Five hypotheses that are known to cause the chronic graft loss were selected (Chen *et al.*, 2013). Those are ER stress (Fonseca *et al.*, 2011; Rickels *et al.*, 2008; Potter *et al.*, 2010; Westermark *et al.*, 2008), islet exhaustion

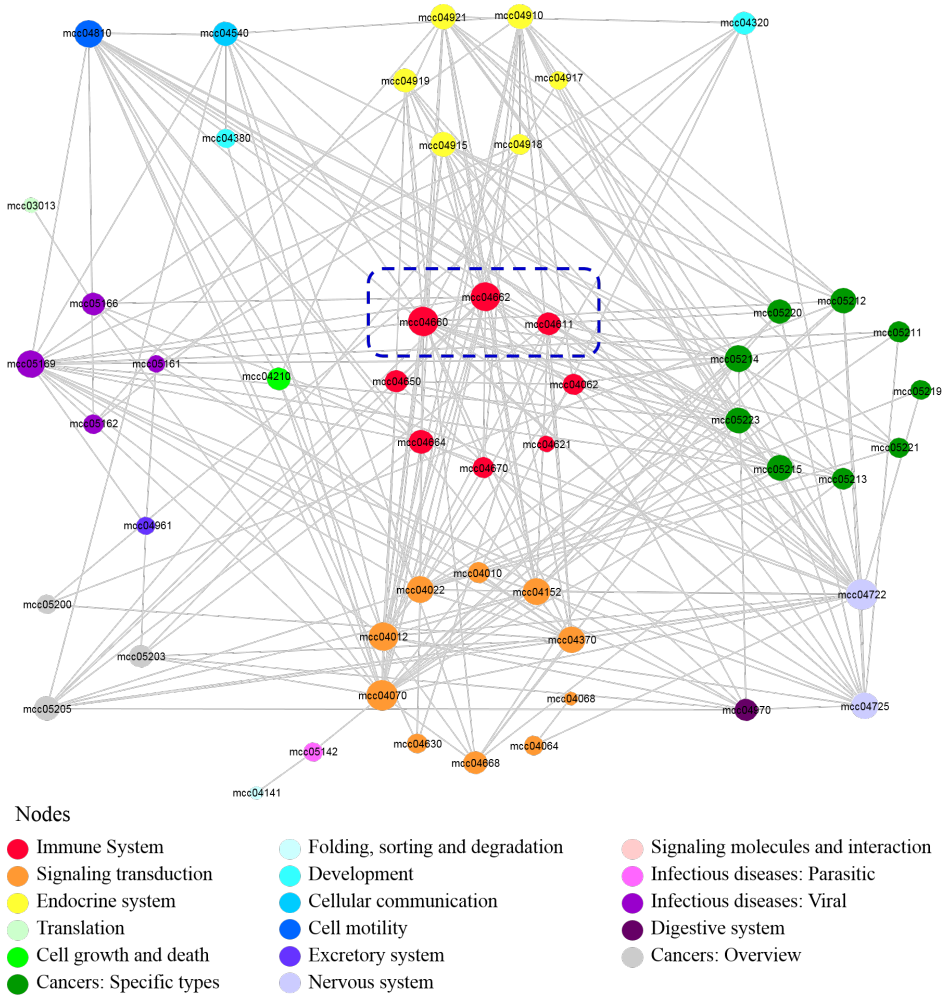


Figure 4.4: Pathway interaction network of GLPAPs

Blue dotted rectangle indicates T cell receptor signaling pathway (mcc04660), B cell receptor signaling pathway (mcc04662), and Platelet activation (mcc04611). The size of the nodes reflects the closeness centrality of each node. The network was visualized by Cytoscape (Shannon *et al.*, 2003)

Table 4.3: The closeness centrality and the degree of GLPAPs in immune system

The average closeness centrality and the average degree of all GLPAPs in the network are 0.4669 and 8.65 respectively and the values were used as the cutoff values to determine if a GLPAP is meaningful in the pathway interaction network. Only T cell receptor signaling pathway, B cell receptor signaling pathway, and Platelet activation satisfied the cutoff values. The pathways are highlighted by underline.

ID	Name	Closeness centrality	Degree
mcc04660	T cell receptor signaling pathway	0.60000000	21
mcc04662	B cell receptor signaling pathway	0.59302326	20
mcc04611	Platelet activation	0.46788991	11
mcc04664	Fc epsilon RI signaling pathway	0.49514563	8
mcc04650	Natural killer cell mediated cytotoxicity	0.45945946	7
mcc04062	Chemokine signaling pathway	0.43589744	7
mcc04670	Leukocyte transendothelial migration	0.43220339	5
mcc04621	NOD-like receptor signaling pathway	0.33774834	1

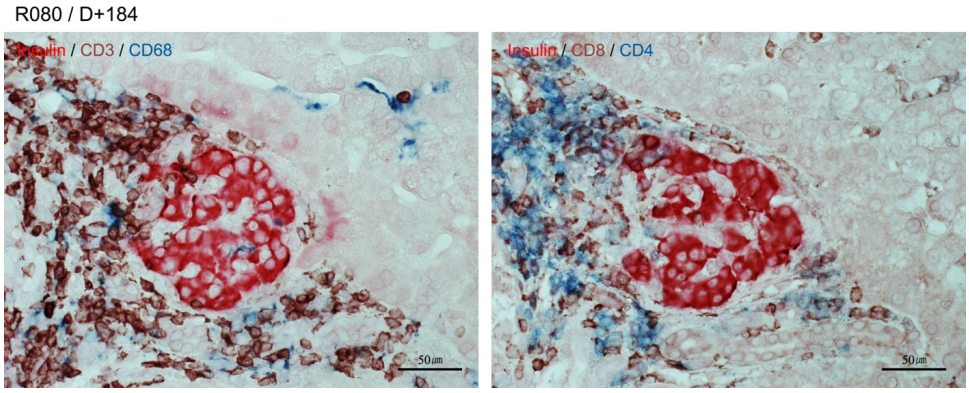


Figure 4.5: Histology of islet xenografts

The islet graft was heavily infiltrated by several types of immune cells in R080. Immune cells largely consisted of $CD3^+$ T cells. Both $CD4^+$ and $CD8^+$ cells infiltrated near the graft. $CD68^+$ cell were also observed. Black arrow indicated intra-graft infiltrating $CD3^+$ T cell.

(Kim and Yoon, 2011), lipotoxicity (Lee *et al.*, 2007; Brown and Goldstein, 1997, 1998; Kakuma *et al.*, 2000), long-term graft rejection, and toxicity of immunosuppressant (Barlow *et al.*, 2013). To evaluate the five hypotheses, I designed and performed a computational experiment. The rationale behind the experiment is that if a hypothesis is the cause of the graft loss and the genes related to the hypothesis are important, the global effects of the genes of the hypothesis should be similar to the gene expression profile that I measured. To measure the global effect of the genes, I used the state of the art network propagation technique (Cowen *et al.*, 2017). The evaluation process is as follows. I collected genes related to five hypotheses as seed genes by the literature search and domain knowledge. Thus, each hypothesis was represented by a set of genes. Next, I made a DEG profile by measuring the expression change of each gene by calculating the log2 fold change between R080 and R051 at t_3 and ranking the genes. At that time, I removed the genes of which the expression

value was smaller than 1 in either R080 or R051 to prevent extremely high or low fold change yielded by the comparison between small numbers. Then I built a protein-protein interaction (PPI) network and mapped the seed genes. The number of nodes and edges in the network are 6,780 and 117,963 respectively. The number of seed genes are ten, nine, eight, ten, and nine for ER stress, islet exhaustion, lipotoxicity, long-term graft rejection, and toxicity of immunosuppressant, respectively. After that, I measured the global effect of the seed genes using network propagation and ranked genes in the PPI network for each hypothesis. Then, I calculated Pearson's correlation between the ranking of the DEG profile and the ranking by the network propagation for each hypothesis. This was to test which of the five hypotheses represented by the seed genes produces gene expression profile similar to the DEG profile I measured. In other words, I tried to see how much the participation of genes in the actual biological process that drives the graft loss coincided with the perturbation in the expression of genes in the given condition. I calculated the correlation coefficients based on the rankings of the network propagation results. Furthermore, I performed random simulations for 1,000 times and calculated empirical p-values to test the significance of the coefficient as shown in the equation below.

$$p^i = \frac{1}{N} \sum_{j=1}^{N} \begin{cases} 1 \text{ if } c_{ij} > c_i^R \\ 0 \text{ otherwise} \end{cases} \tag{4.1}$$

i indicates each hypothesis and it ranges from 1 to 5. p^i indicates the empirical p-value of i-th hypothesis. N is the number of random simulation and it is 1,000. j indicates the j-th random simulation. c_{ij} is the coefficient of j-th random simulation of i-th hypothesis. c_i^R is the reference coefficient of i-th hypothesis. The results are shown in Table 4.4 (a) and I was able to see that the correlation coefficient of long-term graft rejection was the highest and most sig-

nificant. In addition, I carried out the same process for top 100 genes of network propagation results for each hypothesis. As shown in Table 4.4 (b), long-term graft loss was the highest in terms of the coefficient. This result suggests and supports that chronic graft loss reflected by the condition-specific changes in gene expression of R080 was explained the best by long-term graft rejection.

4.3 Discussion

Human islet transplantation is currently the only treatment option which can supplement islet mass for type 1 diabetes patient (McCall and Shapiro, 2012). Although it is successful in short-term, it lacks long-term durability, resulting in most patients transplanted with islets returning to insulin dependency. A pig-to-NHP islet xenotransplantation study was performed to infer the cause of long-term graft rejection. RNA-seq technology was used to quantify the amount of transcript in the samples obtained from whole blood taken at various time points after transplantation. Pathway analysis and pathway interaction network analysis were carried out on the transcriptome data using TRAP and PINTnet, respectively. By performing pathway analyses, 59 activated pathways were identified and they were annotated as graft loss period-related activated pathways (GLPAPs). Furthermore, GLPAPs were categorized to retrieve meaningful information. Indeed, mostly enriched category was revealed as immune system. This highly suggested that cause of graft loss in chronic phase in R080 is due to insufficient immune suppression, i.e. immune-rejection. Subsequently, a pathway interaction network was constructed using GLPAPs as nodes to reveal which pathways played a central role in the given condition and it was found that T cell receptor signaling pathway, B cell signaling pathway and Platelet activation are the most interconnected pathways. This information suggested

Table 4.4: Ranking comparison between network propagation results and differential expression

(a) Pearson's correlation coefficients of each hypothesis. IsletExh, LTGR, and ToxImmDrug indicate islet exhaustion, long-term graft rejection, and toxicity of immunosuppressant in this order. The coefficient of long-term graft rejection was the highest.

(b) Correlation coefficients of ranking comparison for top 100 genes from the network propagation results. Long-term graft rejection showed the highest coefficient.

(a)

Scenario	Coefficient	p-value	Empirical p-value
ERstress	0.031115500	0.001246017	0.048
IsletExh	0.049513322	0.000048063	0.292
Lipotoxicity	0.051612597	0.000022611	0.251
LTGR	0.087461960	0.000000000	0.010
ToxImmDrug	0.050939480	0.000028885	0.275

(b)

Scenario	Coefficient	p-value
ERstress	0.235867693	0.018154182
IsletExh	0.154287565	0.125356981
Lipotoxicity	0.188772704	0.059979769
LTGR	0.556480178	0.000000001
ToxImmDrug	0.536431327	0.000000008

that immune-suppression regimen in the maintenance period should be revised and fortified to overcome the graft loss in chronic phase. These findings were reinforced by hypothesis testing and confirmed by biopsy. However, there are some limitations in the study. First, even though three pathways were suggested to be involved in graft rejection in the chronic phase, direct evidences but T cell infiltration were not found; there was no direct evidence of B cell-mediated or platelet-mediated rejection. Second, because the pathway information for rhesus monkey in KEGG was relatively insufficient, it was not possible to investigate analyze the data in more specific manner. For example, CD40L signaling pathway and IL-6 signaling pathway were not supported by KEGG so I was not able to investigate how the pathways participated in graft rejection. Lastly, I was not able to pinpoint single candidate molecule or a set of molecules responsible for graft rejection because the analyses mainly focused on finding phenotypically relevant pathways.

Chapter 5

Conclusion

The interpretation of complex phenotypes requires the development of novel methods. Because such phenotypes can be represented as networks and there are evidences that networks can explain the complex phenotypes, the development of network analysis is very crucial to reveal the secrets underlying biological process. There are already many existing methods. Thus, making an effective ensemble of those method is as important as the development of new methods. This thesis presented three studies based on network analysis and ensemble of different methods:

1. a machine learning-based approach for identifying DEGs using network information and network propagation
2. a method to construct condition-specific pathway interaction network computing shortest paths on a weighted PPI network
3. a network analysis on time-series xenotransplantation data to reveal the cause of islet graft loss

In the first study, a machine learning-based approach for identifying DEGs using network information and network propagation, MLDEG, is developed. It defines true DEGs and false DEGs by integrating the results of four existing methods and trains a model using network-based features extracted from the DEGs. The goal of MLDEG is to identify DEGs that cannot clearly be detected by the existing methods. Tested on 10 RNA-seq datasets, it was able to rescue the DEGs mentioned to be ground truths in the original papers. Compared to four existing methods, it outperformed the methods. In the second study, a method to construct condition-specific pathway interaction network computing shortest paths on a weighted PPI network, PINTnet, is developed. PINTnet constructs a weighted PPI subnetwork for every pair of pathways by computing shortest paths and measures the activation of pathway interaction using the subnetworks and gene expression data. It ranks the interactions by the activation scores and constructs a pathway interaction network with the interactions that satisfy a cutoff. Three RNA-seq datasets were used to evaluate the performance and the pathways and the interactions relevant to the phenotypes were detected for each dataset. In the last study, pathway analyses including pathway category enrichment and pathway interaction were carried out to find the cause of islet graft loss in porcine islet-transplanted nonhuman primates. The analysis results suggested the activation of T cell signaling pathway as a probable cause and it was confirmed by liver biopsy result. In addition, network propagation was carried out to verify that long-term graft rejection affected the islet graft loss. In conclusion, I developed approaches to carry out transcriptome data analysis from gene level to pathway level using network-based approaches and a machine learning-based approach integrating existing methods. The effective ensemble of the approaches suggests the availability of network analysis in interpreting complex phenotypes. Nevertheless, there are some limitations

in network analysis. First, analysis results can be biased to the genes and interactions that are heavily studied. Second, the effects of hub genes sometimes are too strong to discover subtle but important changes in other genes. Lastly, the analysis results can vary according to the source of network information. Therefore, studying on more robust methods that can overcome the limitations is my future goal.

Bibliography

Akeno, N., Smith, E. P., Stefan, M., Huber, A. K., Zhang, W., Keddache, M., and Tomer, Y. (2011). Ifn-α mediates the development of autoimmunity both by direct tissue toxicity and through immune cell recruitment mechanisms. *The Journal of Immunology*, **186**(8), 4693–4706.

Aksamitiene, E., Kiyatkin, A., and Kholodenko, B. N. (2012). Cross-talk between mitogenic ras/mapk and survival pi3k/akt pathways: a fine balance. *Biochemical Society Transactions*, **40**(1), 139–146.

Ashwell, J. D., King, L. B., and Vacchio, M. S. (1996). Cross-talk between the t cell antigen receptor and the glucocorticoid receptor regulates thymocyte development. *Stem cells*, **14**(5), 490–500.

Assche, F., Aerts, L., and Prins, F. D. (1978). A morphological study of the endocrine pancreas in human pregnancy. *BJOG: An International Journal of Obstetrics & Gynaecology*, **85**(11), 818–820.

Barlow, A. D., Nicholson, M. L., and Herbert, T. P. (2013). Evidence for rapamycin toxicity in pancreatic β-cells and a review of the underlying molecular mechanisms. *Diabetes*, **62**(8), 2674–2682.

Barton, F. B., Rickels, M. R., Alejandro, R., Hering, B. J., Wease, S., Naziruddin, B., Oberholzer, J., Odorico, J. S., Garfinkel, M. R., Levy, M., *et al.* (2012). Improvement in outcomes of clinical islet transplantation: 1999–2010. *Diabetes care*, **35**(7), 1436–1445.

Bernard, V., Young, J., Chanson, P., and Binart, N. (2015). New insights in prolactin: pathological implications. *Nature Reviews Endocrinology*, **11**(5), 265–275.

Bole-Feysot, C., Goffin, V., Edery, M., Binart, N., and Kelly, P. A. (1998). Prolactin (prl) and its receptor: actions, signal transduction pathways and phenotypes observed in prl receptor knockout mice. *Endocrine reviews*, **19**(3), 225–268.

Bosma, N. A., Singla, A. K., Downey, C. M., and Jirik, F. R. (2014). Selumetinib produces a central core of apoptosis in breast cancer bone metastases in mice. *Oncoscience*, **1**(12), 821.

Brooks, M. J., Rajasimha, H. K., Roger, J. E., and Swaroop, A. (2011). Next-generation sequencing facilitates quantitative analysis of wild-type and nrl-/- retinal transcriptomes.

Brown, M. S. and Goldstein, J. L. (1997). The srebp pathway: regulation of cholesterol metabolism by proteolysis of a membrane-bound transcription factor. *Cell*, **89**(3), 331–340.

Brown, M. S. and Goldstein, J. L. (1998). Sterol regulatory element binding proteins (srebps): controllers of lipid synthesis and cellular uptake. *Nutrition reviews*, **56**(suppl_1), S1–S3.

Buijs, J. T., Stayrook, K. R., and Guise, T. A. (2012). The role of tgf-[beta] in bone metastasis: novel therapeutic perspectives. *BoneKEy reports*, **1**(6).

Chen, Y., Shi, H. Y., Stock, S. R., Stern, P. H., and Zhang, M. (2011). Regulation of breast cancer-induced bone lesions by β-catenin protein signaling. *Journal of Biological Chemistry*, **286**(49), 42575–42584.

Chen, Y., Zhang, H., Xiao, X., Jia, Y., Wu, W., Liu, L., Jiang, J., Zhu, B., Meng, X., and Chen, W. (2013). Peripheral blood transcriptome sequencing reveals rejection-relevant genes in long-term heart transplantation. *International journal of cardiology*, **168**(3), 2726–2733.

Cordenonsi, M., Zanconato, F., Azzolin, L., Forcato, M., Rosato, A., Frasson, C., Inui, M., Montagner, M., Parenti, A. R., Poletti, A., *et al.* (2011). The hippo transducer taz confers cancer stem cell-related traits on breast cancer cells. *Cell*, **147**(4), 759–772.

Cowen, L., Ideker, T., Raphael, B. J., and Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, **18**(9), 551.

Deel, M. D., Li, J. J., Crose, L. E., and Linardic, C. M. (2015). A review: molecular aberrations within hippo signaling in bone and soft-tissue sarcomas. *Frontiers in oncology*, **5**.

Deng, X., Zou, W., Xiong, M., Wang, Z., Engelhardt, J. F., Ye, S. Q., Yan, Z., and Qiu, J. (2017). Human parvovirus infection of human airway epithelia induces pyroptotic cell death via inhibiting apoptosis. *Journal of virology*, pages JVI–01533.

Dey, N., Barwick, B. G., Moreno, C. S., Ordanic-Kodani, M., Chen, Z., Oprea-Ilies, G., Tang, W., Catzavelos, C., Kerstann, K. F., Sledge, G. W., *et al.* (2013). Wnt signaling in triple negative breast cancer is associated with metastasis. *BMC cancer*, **13**(1), 1.

Dona, M. S., Prendergast, L. A., Mathivanan, S., Keerthikumar, S., and Salim, A. (2017). Powerful differential expression analysis incorporating network topology for next-generation sequencing data. *Bioinformatics*, **33**(10), 1505–1513.

Donato, M., Xu, Z., Tomoiaga, A., Granneman, J. G., MacKenzie, R. G., Bao, R., Than, N. G., Westfall, P. H., Romero, R., and Draghici, S. (2013). Analysis and correction of crosstalk effects in pathway analysis. *Genome research*, **23**(11), 1885–1893.

Dorr, C., Wu, B., Guan, W., Muthusamy, A., Sanghavi, K., Schladt, D. P., Maltzman, J. S., Scherer, S. E., Brott, M. J., Matas, A. J., *et al.* (2015). Differentially expressed gene transcripts using rna sequencing from the blood of immunosuppressed kidney allograft recipients. *PloS one*, **10**(5), e0125045.

Drachenberg, C. B., Klassen, D. K., Weir, M. R., Wiland, A., Fink, J. C., Bartlett, S. T., Cangro, C. B., Blahut, S., and Papadimitriou, J. C. (1999). Islet cell damage associated with tacrolimus and cyclosporine: Morphological features in pancreas allograft biopsies and clinical correlation1. *Transplantation*, **68**(3), 396–402.

Fang, X., Poulsen, R. R., Rivkees, S. A., and Wendler, C. C. (2016). In utero caffeine exposure induces transgenerational effects on the adult heart. *Scientific reports*, **6**, 34106.

Feldmann, M., Dayan, C., Rapoport, B., and Londei, M. (1992). T cell activation and antigen presentation inhuman thyroid autoimmunity. *Journal of autoimmunity*, **5**, 115–121.

Fisher, R. (1932). Statistical methods for research workers.

Fonseca, S. G., Gromada, J., and Urano, F. (2011). Endoplasmic reticulum stress and pancreatic β-cell death. *Trends in Endocrinology & Metabolism*, **22**(7), 266–274.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., Von Mering, C., *et al.* (2013). String v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, **41**(D1), D808–D815.

Francesconi, M., Remondini, D., Neretti, N., Sedivy, J. M., Cooper, L. N., Verondini, E., Milanesi, L., and Castellani, G. (2008). Reconstructing networks of pathways via significance analysis of their intersections. *BMC bioinformatics*, **9**(4), 1.

Franke, L., Van Bakel, H., Fokkens, L., De Jong, E. D., Egmont-Petersen, M., and Wijmenga, C. (2006). Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *The American Journal of Human Genetics*, **78**(6), 1011–1025.

Glaab, E., Baudot, A., Krasnogor, N., Schneider, R., and Valencia, A. (2012). Enrichnet: network-based gene set enrichment analysis. *Bioinformatics*, **28**(18), i451.

Guardavaccaro, D. and Clevers, H. (2012). Wnt/β-catenin and mapk signaling: allies and enemies in different battlefields. *Sci. Signal.*, **5**(219), pe15–pe15.

Guise, T. A. (2013). Breast cancer bone metastases: it's all about the neighborhood. *Cell*, **154**(5), 957–959.

Guo, X. and Wang, X.-F. (2009). Signaling cross-talk between tgf-β/bmp and other pathways. *Cell research*, **19**(1), 71–88.

Hsu, P.-C., Yang, U.-C., Shih, K.-H., Liu, C.-M., Liu, Y.-L., and Hwu, H.-G. (2008). A protein interaction based model for schizophrenia study. *BMC bioinformatics*, **9**(Suppl 12), S23.

Huang, C., Snider, F., and Cross, J. C. (2009). Prolactin receptor is required for normal glucose homeostasis and modulation of β-cell mass during pregnancy. *Endocrinology*, **150**(4), 1618–1626.

Iozzo, R. V. and Sanderson, R. D. (2011). Proteoglycans in cancer biology, tumour microenvironment and angiogenesis. *Journal of cellular and molecular medicine*, **15**(5), 1013–1031.

Itasaki, N. and Hoppler, S. (2010). Crosstalk between wnt and bone morphogenic protein signaling: a turbulent relationship. *Developmental Dynamics*, **239**(1), 16–33.

Jamieson, C. A. and Yamamoto, K. R. (2000). Crosstalk pathway for inhibition of glucocorticoid-induced apoptosis by t cell receptor signaling. *Proceedings of the National Academy of Sciences*, **97**(13), 7319–7324.

Jang, G.-B., Kim, J.-Y., Cho, S.-D., Park, K.-S., Jung, J.-Y., Lee, H.-Y., Hong, I.-S., and Nam, J.-S. (2015). Blockade of wnt/β-catenin signaling suppresses breast cancer metastasis by inhibiting csc-like phenotype. *Scientific reports*, **5**.

Jo, K., Kwon, H.-B., and Kim, S. (2014). Time-series rna-seq analysis package (trap) and its application to the analysis of rice, oryza sativa l. ssp. japonica, upon drought stress. *Methods*, **67**(3), 364–372.

Kakuma, T., Lee, Y., Higa, M., Wang, Z.-w., Pan, W., Shimomura, I., and Unger, R. H. (2000). Leptin, troglitazone, and the expression of sterol regulatory element binding proteins in liver and pancreatic islets. *Proceedings of the National Academy of Sciences*, **97**(15), 8536–8541.

Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **28**(1), 27–30.

Karnoub, A. E. and Weinberg, R. A. (2008). Ras oncogenes: split personalities. *Nature reviews Molecular cell biology*, **9**(7), 517–531.

Kawashima, A., Tanigawa, K., Akama, T., Yoshihara, A., Ishii, N., and Suzuki, K. (2011). Innate immune activation and thyroid autoimmunity. *The Journal of Clinical Endocrinology & Metabolism*, **96**(12), 3661–3671.

Kim, H., Toyofuku, Y., Lynn, F. C., Chak, E., Uchida, T., Mizukami, H., Fujitani, Y., Kawamori, R., Miyatsuka, T., Kosaka, Y., *et al.* (2010). Serotonin regulates pancreatic beta cell mass during pregnancy. *Nature medicine*, **16**(7), 804–808.

Kim, J., Okamoto, H., Huang, Z., Anguiano, G., Chen, S., Liu, Q., Cavino, K., Xin, Y., Na, E., Hamid, R., *et al.* (2017). Amino acid transporter slc38a5 controls glucagon receptor inhibition-induced pancreatic α cell hyperplasia in mice. *Cell metabolism*, **25**(6), 1348–1361.

Kim, J.-W. and Yoon, K.-H. (2011). Glucolipotoxicity in pancreatic β-cells. *Diabetes & metabolism journal*, **35**(5), 444–450.

Kim, S.-Y. and Volsky, D. J. (2005). Page: parametric analysis of gene set enrichment. *BMC bioinformatics*, **6**(1), 144.

Kotsiantis, S. B., Zaharakis, I., and Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, **160**, 3–24.

Lau, J., Henriksnäs, J., Svensson, J., and Carlsson, P.-O. (2009). Oxygenation of islets and its role in transplantation. *Current opinion in organ transplantation*, **14**(6), 688–693.

Lee, Y., Ravazzola, M., Park, B.-H., Bashmakov, Y. K., Orci, L., and Unger, R. H. (2007). Metabolic mechanisms of failure of intraportally transplanted pancreatic β-cells in rats: role of lipotoxicity and prevention by leptin. *Diabetes*, **56**(9), 2295–2301.

Leitão, C. B., Bernetti, K., Tharavanij, T., Cure, P., Lauriola, V., Berggren, P.-O., Ricordi, C., and Alejandro, R. (2010). Lipotoxicity and decreased islet graft survival. *Diabetes Care*, **33**(3), 658–660.

Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M., Haag, J. D., Gould, M. N., Stewart, R. M., and Kendziorski, C. (2013). Ebseq: an empirical bayes hierarchical model for inference in rna-seq experiments. *Bioinformatics*, **29**(8), 1035–1043.

Lever, J., Krzywinski, M., and Altman, N. (2016). Points of significance: Logistic regression. *Nature Methods*, **13**(7), 541–542.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, **15**(12), 550.

Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D., and Woolf, P. J. (2009). Gage: generally applicable gene set enrichment for pathway analysis. *BMC bioinformatics*, **10**(1), 161.

Malik, S., Sadhu, S., Elesela, S., Pandey, R. P., Chawla, A. S., Sharma, D., Panda, L., Rathore, D., Ghosh, B., Ahuja, V., *et al.* (2017). Transcription factor foxo1 is essential for il-9 induction in t helper cells. *Nature communications*, **8**(1), 815.

Marigliano, M., Bertera, S., Grupillo, M., Trucco, M., and Bottino, R. (2011). Pig-to-nonhuman primates pancreatic islet xenotransplantation: an overview. *Current diabetes reports*, **11**(5), 402.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, **18**(9), 1509–1517.

Matsumoto, M., Sudo, T., Maruyama, M., Osada, H., and Tsujimoto, M. (2000). Activation of p38 mitogen-activated protein kinase is crucial in osteoclastogenesis induced by tumor necrosis factor. *FEBS letters*, **486**(1), 23–28.

McCall, M. and Shapiro, A. J. (2012). Update on islet transplantation. *Cold Spring Harbor perspectives in medicine*, **2**(7), a007823.

Medina, I., Montaner, D., Bonifaci, N., Pujana, M. A., Carbonell, J., Tarraga, J., Al-Shahrour, F., and Dopazo, J. (2009). Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies. *Nucleic acids research*, **37**(suppl 2), W340–W344.

Merrill, S. J. and Mu, Y. (2015). Thyroid autoimmunity as a window to autoimmunity: an explanation for sex differences in the prevalence of thyroid autoimmunity. *Journal of theoretical biology*, **375**, 95–100.

Moon, J. H., Lim, S., Jo, K., Lee, S., Seo, S., and Kim, S. (2017). Pintnet: construction of condition-specific pathway interaction network by computing shortest paths on weighted ppi. *BMC systems biology*, **11**(2), 15.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, **5**(7), 621.

Nam, D., Kim, J., Kim, S.-Y., and Kim, S. (2010). Gsa-snp: a general approach for gene set analysis of polymorphisms. *Nucleic acids research*, page gkq428.

Nam, S., Chang, H. R., Kim, K.-T., Kook, M.-C., Hong, D., Kwon, C., Jung, H. R., Park, H. S., Powis, G., Liang, H., *et al.* (2014). Pathome: an algorithm for accurately detecting differentially expressed subpathways. *Oncogene*, **33**(41), 4941.

Negi, S., Park, S. H., Jetha, A., Aikin, R., Tremblay, M., and Paraskevas, S. (2012). Evidence of endoplasmic reticulum stress mediating cell death in transplanted human islets. *Cell transplantation*, **21**(5), 889–900.

Nielsen, J. H. (1982). Effects of growth hormone, prolactin, and placental lactogen on insulin content and release, and deoxyribonucleic acid synthesis in cultured pancreatic islets. *Endocrinology*, **110**(2), 600–606.

Oti, M., Snel, B., Huynen, M. A., and Brunner, H. G. (2006). Predicting disease genes using protein–protein interactions. *Journal of medical genetics*, **43**(8), 691–698.

Pan, D. (2010). The hippo signaling pathway in development and cancer. *Developmental cell*, **19**(4), 491–505.

Parsons, J. A., Brelje, T. C., and Sorenson, R. L. (1992). Adaptation of islets of langerhans to pregnancy: increased islet cell proliferation and insulin secretion correlates with the onset of placental lactogen secretion. *Endocrinology*, **130**(3), 1459–1466.

Pataki, C. A., Couchman, J. R., and Brábek, J. (2015). Wnt signaling cascades and the roles of syndecan proteoglycans. *Journal of Histochemistry & Cytochemistry*, **63**(7), 465–480.

Paulmann, N., Grohmann, M., Voigt, J.-P., Bert, B., Vowinckel, J., Bader, M., Skelin, M., Jevšek, M., Fink, H., Rupnik, M., *et al.* (2009). Intracellular serotonin modulates insulin secretion from pancreatic β-cells by protein serotonylation. *PLoS Biol*, **7**(10), e1000229.

Potlukova, E. and Limanova, Z. (2006). [the role of complement in autoimmune thyroid disorders]. *Casopis lekaru ceskych*, **146**(3), 210–214.

Potter, K., Abedini, A., Marek, P., Klimek, A., Butterworth, S., Driscoll, M., Baker, R., Nilsson, M., Warnock, G., Oberholzer, J., *et al.* (2010). Islet amyloid deposition limits the viability of human islet grafts but not porcine islet grafts. *Proceedings of the National Academy of Sciences*, **107**(9), 4305–4310.

Pugliese, A., Reijonen, H. K., Nepom, J., and Burke III, G. W. (2011). Recurrence of autoimmunity in pancreas transplant patients: research update. *Diabetes management (London, England)*, **1**(2), 229.

Ramanan, V. K., Shen, L., Moore, J. H., and Saykin, A. J. (2012). Pathway

analysis of genomic data: concepts, methods, and prospects for future development. *TRENDS in Genetics*, **28**(7), 323–332.

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C. E., Socci, N. D., and Betel, D. (2013). Comprehensive evaluation of differential gene expression analysis methods for rna-seq data. *Genome biology*, **14**(9), 1.

Ravà, M., D'andrea, A., Doni, M., Kress, T. R., Ostuni, R., Bianchi, V., Morelli, M. J., Collino, A., Ghisletti, S., Nicoli, P., *et al.* (2017). Mutual epithelium-macrophage dependency in liver carcinogenesis mediated by st18. *Hepatology*, **65**(5), 1708–1719.

Rawlings, J. S., Rosler, K. M., and Harrison, D. A. (2004). The jak/stat signaling pathway. *Journal of cell science*, **117**(8), 1281–1283.

Rickels, M. R., Collins, H. W., and Naji, A. (2008). Amyloid and transplanted islets. *The New England journal of medicine*, **359**(25), 2729.

Rieck, S. and Kaestner, K. H. (2010). Expansion of β-cell mass in response to pregnancy. *Trends in Endocrinology & Metabolism*, **21**(3), 151–158.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, **43**(7), e47–e47.

Rivals, I., Personnaz, L., Taing, L., and Potier, M.-C. (2007). Enrichment or depletion of a go category within a class of genes: which test? *Bioinformatics*, **23**(4), 401–407.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a biocon-

ductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1), 139–140.

Ryan, E. A., Paty, B. W., Senior, P. A., Bigam, D., Alfadhli, E., Kneteman, N. M., Lakey, J. R., and Shapiro, A. J. (2005). Five-year follow-up after clinical islet transplantation. *Diabetes*, **54**(7), 2060–2069.

Sathya, R. and Abraham, A. (2013). Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence*, **2**(2), 34–38.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, **13**(11), 2498–2504.

Shapiro, A. J., Lakey, J. R., Ryan, E. A., Korbutt, G. S., Toth, E., Warnock, G. L., Kneteman, N. M., and Rajotte, R. V. (2000). Islet transplantation in seven patients with type 1 diabetes mellitus using a glucocorticoid-free immunosuppressive regimen. *New England Journal of Medicine*, **343**(4), 230–238.

Sheikh, B., Bechtel-Walz, W., Lucci, J., Karpiuk, O., Hild, I., Hartleben, B., Vornweg, J., Helmstädter, M., Sahyoun, A., Bhardwaj, V., *et al.* (2015). Mof maintains transcriptional programs regulating cellular stress response. *Oncogene*.

Shin, H., Shannon, C. P., Fishbane, N., Ruan, J., Zhou, M., Balshaw, R., Wilson-McManus, J. E., Ng, R. T., McManus, B. M., Tebbutt, S. J., *et al.* (2014). Variation in rna-seq transcriptome profiles of peripheral whole blood

from healthy individuals with and without globin depletion. *PLoS One*, **9**(3), e91041.

Shin, J., Kim, J., Kim, J., Min, B., Kim, Y., Kim, H., Jang, J., Yoon, I., Kang, H., Kim, J., *et al.* (2015). Long-term control of diabetes in immunosuppressed nonhuman primates (nhp) by the transplantation of adult porcine islets. *American Journal of Transplantation*, **15**(11), 2837–2850.

Simeone, O. *et al.* (2018). A brief introduction to machine learning for engineers. *Foundations and Trends® in Signal Processing*, **12**(3-4), 200–431.

Sorenson, R. L., Brelje, T. C., Hegre, O. D., Marshall, S., Anaya, P., and Sheridan, J. D. (1987). Prolactin (in vitro) decreases the glucose stimulation threshold, enhances insulin secretion, and increases dye coupling among islet b cells*. *Endocrinology*, **121**(4), 1447–1453.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., *et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, **102**(43), 15545–15550.

Sukhtankar, D., Okun, A., Chandramouli, A., Nelson, M. A., Vanderah, T. W., Cress, A. E., Porreca, F., and King, T. (2011). Inhibition of p38-mapk signaling pathway attenuates breast cancer induced bone pain and disease progression in a murine model of cancer-induced bone pain. *Molecular pain*, **7**(1), 1.

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N. T., Roth, A., Bork, P., *et al.* (2016). The string

database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research*, page gkw937.

Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-s., Kim, C. J., Kusanovic, J. P., and Romero, R. (2008). A novel signaling pathway impact analysis. *Bioinformatics*, **25**(1), 75–82.

Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-s., Kim, C. J., Kusanovic, J. P., and Romero, R. (2009). A novel signaling pathway impact analysis. *Bioinformatics*, **25**(1), 75–82.

Tegge, A. N., Sharp, N., and Murali, T. (2016). Xtalk: a path-based approach for identifying crosstalk between signaling pathways. *Bioinformatics*, **32**(2), 242–251.

Toubi, E. and Shoenfeld, Y. (2004). Toll-like receptors and their role in the development of autoimmune diseases. *Autoimmunity*, **37**(3), 183–188.

Vadlakonda, L., Pasupuleti, M., and Pallu, R. (2014). Role of pi3k-akt-mtor and wnt signaling pathways in transition of g1-s phase of cell cycle in cancer cells. *Targeting PI3K/mTOR signaling in cancer*, page 87.

Velasco, C. R., Colliec-Jouault, S., Redini, F., Heymann, D., and Padrines, M. (2010). Proteoglycans on bone tumor development. *Drug discovery today*, **15**(13), 553–560.

Wang, D., Kon, N., Lasso, G., Jiang, L., Leng, W., Zhu, W.-G., Qin, J., Honig, B., and Gu, W. (2016a). Acetylation-regulated interaction between p53 and set reveals a widespread regulatory mode. *Nature*, **538**(7623), 118.

Wang, D., Li, J.-R., Zhang, Y.-H., Chen, L., Huang, T., and Cai, Y.-D. (2018).

Identification of differentially expressed genes between original breast cancer and xenograft using machine learning algorithms. *Genes*, **9**(3), 155.

Wang, J., Rouse, C., Jasper, J. S., and Pendergast, A. M. (2016b). Abl kinases promote breast cancer osteolytic metastasis by modulating tumor-bone interactions through taz and stat5 signaling. *Sci. Signal.*, **9**(413), ra12–ra12.

Wang, S. H. and Baker Jr, J. R. (2007). The role of apoptosis in thyroid autoimmunity. *Thyroid*, **17**(10), 975–979.

Wang, X., Dalkic, E., Wu, M., and Chan, C. (2008). Gene module level analysis: identification to networks and dynamics. *Current opinion in biotechnology*, **19**(5), 482–491.

Westermark, G. T., Westermark, P., Berne, C., and Korsgren, O. (2008). Widespread amyloid deposition in transplanted human pancreatic islets. *New England Journal of Medicine*, **359**(9), 977–979.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wu, B. (2005). Differential gene expression detection and sample classification using penalized linear regression models. *Bioinformatics*, **22**(4), 472–476.

Xing, L., Dai, Z., Jabbari, A., Cerise, J. E., Higgins, C. A., Gong, W., De Jong, A., Harel, S., DeStefano, G. M., Rothman, L., *et al.* (2014). Alopecia areata is driven by cytotoxic t lymphocytes and is reversed by jak inhibition. *Nature medicine*, **20**(9), 1043.

Zhang, W., Johnson, N., Wu, B., and Kuang, R. (2012). Signed network propagation for detecting differential gene expressions and dna copy number vari-

ations. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pages 337–344. ACM.

Zheng, W., O'Hear, C. E., Alli, R., Basham, J. H., Abdelsamed, H. A., Palmer, L. E., Jones, L. L., Youngblood, B., and Geiger, T. L. (2018). Pi3k orchestration of the in vivo persistence of chimeric antigen receptor-modified t cells. *Leukemia*, **32**(5), 1157–1167.

Zheng, X., Wang, X., Ma, Z., Sunkari, V. G., Botusan, I., Takeda, T., Björklund, A., Inoue, M., Catrina, S., Brismar, K., *et al.* (2012). Acute hypoxia induces apoptosis of pancreatic β-cell by activation of the unfolded protein response and upregulation of chop. *Cell death & disease*, **3**(6), e322.

Zhu, Y., Lyapichev, K., Lee, D., Motti, D., Ferraro, N., Zhang, Y., Yahn, S., Soderblom, C., Zha, J., Bethea, J., *et al.* (2017). Macrophage transcriptional profile identifies lipid catabolic pathways that can be therapeutically targeted after spinal cord injury. *Journal of Neuroscience*, pages 2751–16.

Zwerina, J., Hayer, S., Redlich, K., Bobacz, K., Kollias, G., Smolen, J. S., and Schett, G. (2006). Activation of p38 mapk is a key step in tumor necrosis factor–mediated inflammatory bone destruction. *Arthritis & Rheumatism*, **54**(2), 463–472.

초록

전사 과정에서의 생물학적 프로세스에 대한 이해를 높이는 데 사용되는 전사체 데이터의 분석은 차별 발현 유전자를 찾아내는 것에서부터 표현형에 연관된 패스웨이 증폭 분석까지의 일련의 단계를 포함한다. 각 단계마다, 넘어야 할 장애물들이 존재하며 이를 극복하기 위한 새로운 생물정보학 기술의 개발은 필수적이다. 예를 들어, 생명체의 복잡한 특성은 유전자 또는 패스웨이가 노드, 그 개체 사이의 상호작용이 엣지인 네트워크로 나타낼 수 있다. 이 때, 네트워크 분석 기법은 전사체 데이터와 표현형 간의 숨겨진 연관성을 찾는 데 중요한 역할을 할 수 있다. 한 편, 네트워크 전파는 네트워크에서 노드의 영향력을 측정하는 기술로 주목받고 있으며 새로운 생물학적 발견에 기여하는 등, 생물학 및 의학 분야의 많은 연구에서 그 유용성을 입증하였다. 본 논문에서는 이러한 기계 학습, 네트워크 정보 및 네트워크 전파를 이용한 전사체 데이터 분석에 관한 연구에 대해 다룬다.

첫 번째 연구에서는, 네트워크 정보와 네트워크 전파를 이용하여 차별 발현 유전자를 식별하는 기계 학습 접근법(MLDEG)에 관한 연구를 다룬다. 차별 발현 유전자 분석은 생물학 연구에서 새로운 생물학적 지식의 발견에 중요한 역할을 하고 있으나 이를 위한 기존의 분석 도구들이 도출하는 결과는 각기 다르다. 본 연구에서는 네트워크 정보 및 네트워크 전파 결과를 활용하는 모델을 구축하여 이러한 문제를 해결하였다. 본 연구의 목표는 차별 발현 유전자 및 비차별 발현 유전자로서 가장 가능성이 있는 유전자를 선정하여 네트워크 기반 특징을 추출하고 이 특징을 바탕으로 모델을 학습하여 차별 발현 유전자를 분류하는 것이다. 열 개의 RNA-seq 데이터를 이용하여 검증한 결과, 기존의 분석 도구들보다 우수한 성능을 보임을 확인하였다.

두 번째 연구에서는 단백질 상호 작용 네트워크상의 최단 경로를 계산하여 특정 실험 조건하에서 패스웨이 상호 작용 네트워크를 구축할 수 있는 패스웨이 상호

작용 네트워크 구축 방법(PINTnet)에 대한 내용을 다룬다. 기존의 방법들은 유전자 사이의 관계를 고려하지 않고 패스웨이를 단순히 유전자의 집합으로만 다루는 문제를 가지고 있다. 본 연구에서는 유전자 사이의 관계를 고려하여 각 패스웨이 쌍에 매핑된 단백질 상호작용 네트워크에서 최단 경로를 계산하고, 이를 통해 만들어진 서브네트워크에서 근접중심성과 유전자 발현량의 곱을 바탕으로 패스웨이 상호작용의 활성화 상태를 측정함으로 문제를 해결하였다. 세 개의 RNA-seq 데이터를 이용하여 PINTnet의 성능을 평가한 결과, 각 데이터의 원 논문에서 주장한 결과를 성공적으로 재현함을 확인하였다.

마지막 연구는 만성 췌도 이식편 소실의 원인을 밝히기 위한 이종장기이식 데이터 분석에 관한 내용을 다룬다. 만성 단계에서의 이식편 소실의 기작을 밝히기 위해, PINTnet을 사용하여 돼지 췌도가 이식된 원숭이의 RNA-seq 데이터를 분석하였고 T 세포 수용체 신호 전달 패스웨이(T cell receptor signalling pathway)가 활성화 되었음을 확인하였다. 해당 원숭이의 간 샘플을 생검하여 $CD3^+$ T 세포가 이식된 췌도에 침투하였음을 확인함으로써 분석 결과가 실제 결과와 일치함을 확인하였다. 한편, 네트워크 전파를 이용하여 다섯 가지 거부 반응 시나리오를 검증하였고 T 세포로 인한 거부반응이 가장 가능성이 높음을 확인하였다.

결론적으로, 본 논문에서는 다양한 전사체 데이터 분석을 수행함에 있어서 네트워크 정보, 네트워크 특성 및 네트워크 전파를 이용한 네트워크 분석 및 기계학습 기법이 유용함을 보였다.

주요어: 단백질 상호작용, 최단거리, 네트워크 전파, 차별 발현 유전자, 이종장기이식, 만성 이식편 소실

학번: 2012-30906